

# Quantitative Text Analysis in Arabic

## التحليل الكمي للنص المكتوب في اللغة العربية

Richard Nielsen

MIT

March 31 - April 4, 2019

Cairo University

# Introduction

Who am I?

Who are you?

Research experience?

Statistics training?

Text analysis experience?

Programming experience?

Have you ever used R?

# Introduction

## Learning Text Analysis the Hard Way

- What is the hard way?
- Why learn the hard way?

Code: <http://www.mit.edu/~rnielsen/arabicTextWorkshop.zip>

Slides: [http://www.mit.edu/~rnielsen/arabic\\_text\\_slides.pdf](http://www.mit.edu/~rnielsen/arabic_text_slides.pdf)

<http://www.mit.edu/~rnielsen/> > “Helpful Stuff” (at the bottom)

# Introduction

In this session:

- Examples of text analysis
- Principles of text analysis research

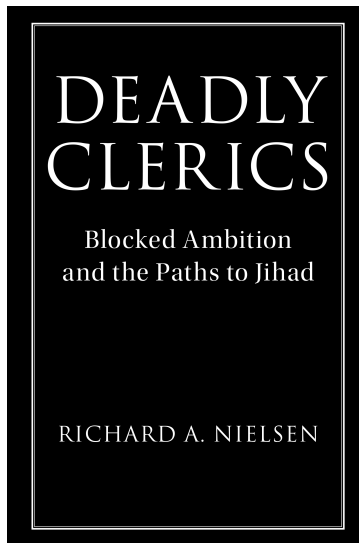


# Introduction

What is quantitative text analysis?

How do I use it in my research?

# Introduction





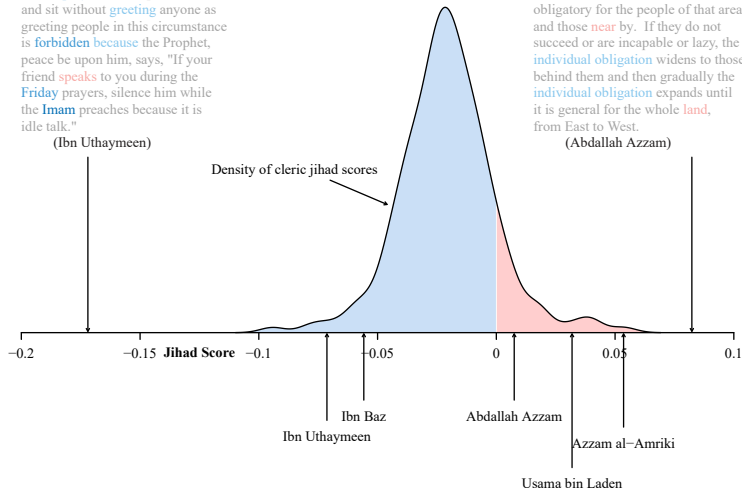
# Introduction

If a **person** arrives while the **Imam** is preaching at **Friday** prayers, he should **pray** two brief prostrations and sit without **greeting** anyone as greeting people in this circumstance is **forbidden because** the Prophet, peace be upon him, says, "If your friend **speaks** to you during the **Friday** prayers, silence him while the **Imam** preaches because it is idle talk."

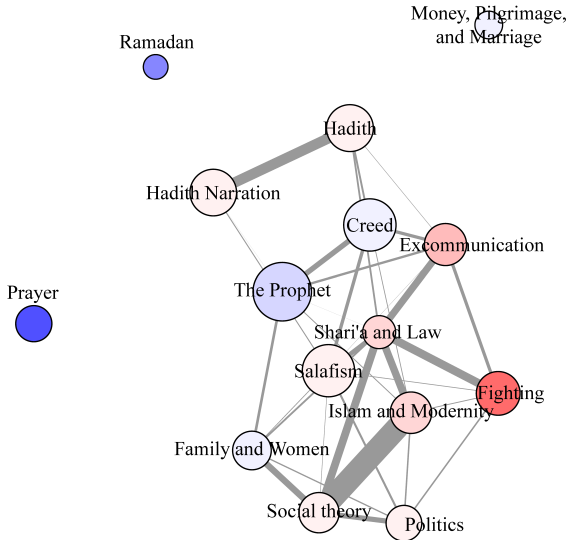
(Ibn Uthaymeen)

Ruling on **Fighting** Now in **Palestine** and **Afghanistan**. The foregoing has clarified that if an inch of Muslim lands are attacked, then **Jihad** is obligatory for the people of that area, and those **near** by. If they do not succeed or are incapable or lazy, the **individual obligation** widens to those behind them and then gradually the **individual obligation** expands until it is general for the whole **land**, from East to West.

(Abdallah Azzam)



# Introduction



# Introduction

## Women's Authority in Patriarchal Social Movements: The Case of Female Salafi Preachers

Richard A. Nielsen\*

February 26, 2019

Forthcoming: *American Journal of Political Science*

### Abstract

How do women gain authority in the public sphere, especially in contexts where patriarchal norms are prevalent? I argue that the leaders of patriarchal social movements face pragmatic incentives to expand women's authority roles when seeking new movement members. Women authorities help patriarchal movements by making persuasive, identity-based arguments in favor of patriarchy that men cannot, and by reaching new audiences that men cannot. I support this argument by examining the rise of online female preachers in the Islamist Salafi movement, using interviews, Twitter analysis, and automated text analysis of 21,000 texts by 172 men and 43 women on the Salafi-oriented website [saaid.net](http://saaid.net). To show the theory's generality, I also apply it to the contemporary white nationalist movement in the United States. The findings illustrate how movements that aggressively enforce traditional gender roles for participants can nevertheless increase female authority for pragmatic political reasons.



saaaid.net

# صيد الفوائد

صيد الفوائد



البحث

ساهم معنا

اتصل بنا

أخبر صديقك

العروض الدعوية

المكتبة

الرئيسية

صيد الفوائد الأنشطة الدعوية

صيد الفوائد جديد المكتبة

• 53 قاعدة لحياة أفضل

• قراءة نقدية لكتاب (ما بعد

الصحوة) للأساذ عبد الله

الغلامي ...!

• قلوب نهوى الحطاء

• همسة في أذن زوجين

• تعريفات في علم الاجتماع

الديني

• أحوال الأمم وسنن الله تعالى

في الاجتماع البشري

صيد الفوائد مواقع اسلامية

الآن

الموقع باللغة الإنجليزية



تغريدات بواسطة @saaaidnet1

موقع صيد الفوائد @saaaidnet1

بين يدي الأمير!

أحمد بن عبد المحسن الحساف

@ahmalassaf

مشاهدة على تويتر

إدراج

مختارات

صيد الفوائد

الفرق المنهجية بين أهل السنة وأهل البدعة

بين يدي الأمير!

عشرة أبيات في معنى

النقرآن : إعجاز علمي

التدوين : الكلمة حين تنتصر!

صيد الفوائد الرئيسية

• اعرف نبيك

• العلماء وطبئة العلم

• أفكار دعوية

• مكتبة صيد الفوائد

• الأنشطة الدعوية

• زاد الداعية

• زاد الخطيب

• العروض الدعوية

• للنساء فقط

• ملتقى الداعيات

• رسائل دعوية

• انفلاشات - القصص

• مقالات - تغريدات

• واحة الأدب

• منوعات - مختارات

• المثل والنحل

• الطبيب الداعية

• بحوث علمية

• تربية الأبناء

• سداة الشريعة



saaaid.net

# صيد الفوائد

صيد الفوائد



البحث

ساهم معنا

اتصل بنا

أخبر صديقك

العروض الدعوية

المكتبة

الرئيسية

صيد الفوائد الأنشطة الدعوية

صيد الفوائد جديد المكتبة

- 53 قاعدة لحياة أفضل
- قراءة نقدية لكتاب (ما بعد الصحوة) للأساذ عبد الله العظامي ...!
- قلوب نهوى الحطاء
- همسة في أذن زوجين
- تعريفات في علم الاجتماع الدني
- أحوال الأمم وسنن الله تعالى في الاجتماع البشري

صيد الفوائد مواقع اسلامية

الآن

الموقع باللغة الإنجليزية



## المعلم الداعية



## مكتبة صيد الفوائد

تغريدات بواسطة @saaaidnet1



@saaaidnet1 موقع صيد الفوائد

بين يدي الأمير!

أحمد بن عبد المحسن الحساف

@ahmalassaf



مشاهدة على تويتر

إدراج

صيد الفوائد مختارات

الفرق المتهجئة بين أهل السنة وأهل البدعة

بين يدي الأمير!

عشرة أبيات في معنى

النقرآن : إعجاز علمي

التدوين : الكلمة حين تنتصر!

صيد الفوائد الرئيسية

- اعرف نبيك
- العلماء وطبقة العلم
- أفكار دعوية
- مكتبة صيد الفوائد
- الأنشطة الدعوية
- زاد الداعية
- زاد الخطيب
- العروض الدعوية
- للنساء فقط
- ملتقى الداعيات
- رسائل دعوية
- انفلاشات - القصص
- مقالات - تغريدات
- واحة الأدب
- منوعات - مختارات
- المثل والنحل
- الطبيب الداعية
- بحوث علمية
- تربية الأبناء
- سداة الشريعة





saaid.net

# صيد الفوائد

صيد الفوائد



البحث

ساهم معنا

اتصل بنا

أقرب مدقق

العروض الدعوية

المكتبة

الرئيسية

صيد الفوائد ملتقى طلبة العلم

- رابط 1
- رابط 2
- رابط 3
- رابط 4

صيد الفوائد مواقع اسلامية



## ملتقى العلماء وطلبة العلم

<p>موقع الشيخ محمد العثيمين رحمه الله <b>كتبه</b></p>	<p>موقع الشيخ عبدالعزيز بن باز رحمه الله <b>كتبه</b></p>
<p>موقع الشيخ صالح بن فوزان الفوزان <b>كتبه</b></p>	<p>موقع الشيخ عبدالله الجبرين <b>كتبه</b></p>
<p>موقع الشيخ عبدالرحمن السعدي <b>كتبه</b></p>	<p>الشيخ محمد ناصر الدين الألباني</p>
	<p>الشيخ عبدالرحمن بن ناصر البراك</p>
<p>الشيخ عبد العزيز بن محمد السلمان عبد الله بن جار الله آل جار الله رحمه</p>	<p>الشيخ فيصل بن عبدالعزيز آل مبارك العلامة عبد الرحمن بن محمد بن</p>

صيد الفوائد الرئيسية

- احرف نبيك
- العلماء وطلبة العلم
- أفكار دعوية
- مكتبة صيد الفوائد
- الأنشطة الدعوية
- زاد الداعية
- زاد الخطيب
- العروض الدعوية
- للنساء فقط
- منتقى الداعيات
- رسائل دعوية
- الفلاشات - القصص
- مقالات - تعرييدات
- واحة الأدب
- منوعات - مختارات
- المثل والنحل
- الطبيب الداعية
- بحوث علمية
- تربية الأبناء
- سيادة الشريعة



saaid.net

# صيد الفوائد

صيد الفوائد



البحث

ساهم معنا

اتصل بنا

أقرب مدقق

العروض الدعوية

المكتبة

الرئيسية

صيد الفوائد ملتقى طلبة العلم

- رابط 1
- رابط 2
- رابط 3
- رابط 4

صيد الفوائد مواقع اسلامية



## ملتقى العلماء وطلبة العلم

موقع الشيخ محمد العثيمين رحمه الله كتبه	موقع الشيخ عبدالعزيز بن باز رحمه الله كتبه
موقع الشيخ صالح بن فوزان الفوزان كتبه	موقع الشيخ عبدالله الجبرين كتبه
موقع الشيخ عبدالرحمن السعدي كتبه	الشيخ محمد ناصر الدين الألباني
	الشيخ عبدالرحمن بن ناصر البراك
الشيخ عبد العزيز بن محمد السلمان عبد الله بن جار الله آل جار الله رحمه	الشيخ فيصل بن عبدالعزيز آل مبارك العلامة عبد الرحمن بن محمد بن

صيد الفوائد الرئيسية

- احرف نبيك
- العلماء وطلبة العلم
- أفكار دعوية
- مكتبة صيد الفوائد
- الأنشطة الدعوية
- زاد الداعية
- زاد الخطيب
- العروض الدعوية
- للنساء فقط
- ملتقى الداعيات
- رسائل دعوية
- الفلاشات - القصص
- مقالات - تعرييدات
- واحة الأدب
- منوعات - مختارات
- المثل والنحل
- الطبيب الداعية
- بحوث علمية
- تربية الأبناء
- سيادة الشريعة



saaaid.net

# صيد الفوائد

صيد الفوائد



البحث

ساهم معنا

اتصل بنا

أقرب مديقل

العروض الدعوية

المكتبة

الرئيسية

## القصص المتنوعة لاداعات

- متنقي اداعات
- د. أميمة الجلاهية
- أفراح الحميضي
- رقية المحارب
- د. بنت الرسالة
- أمها الجريس
- فاطمة البطاح
- ندى اليوسفي
- خيرية الحارثي
- د. نهى قاطرجي
- فوزية الخليوي
- شيماء علي
- إكرام الأزدي
- نور الجندي
- عطاء أم معاذ
- د. جواهر آل الشيخ
- صباح الضامن
- سحر لثان
- صاحبة قلم
- لثان شريف

## ملتنقي الداعيات

## الدكتورة أميمة بنت أحمد الجلاهية

## الدكتورة أفراح بنت علي الحميضي

## الدكتورة رقية بنت محمد المحارب

## الدكتورة فوزية بنت محمد

## القصص المتنوعة لاداعات

- اعرف نبيك
- العلماء وطلبة العلم
- أفكار دعوية
- مكتبة صيد الفوائد
- الأنشطة الدعوية
- زاد الداعية
- زاد الخطيب
- العروض الدعوية
- للنساء فقط
- متنقي اداعات
- رسائل دعوية
- الفلاشات - القصص
- مقالات - تفريعات
- واحة الأدب
- منوعات - مختارات
- المثل والنحل
- الطيب الدعوية
- بحوث علمية
- تربية الأبناء
- سعادة الشريعة



saaaid.net

# صيد الفوائد

صيد الفوائد



البحث

ساهم معنا

اتصل بنا

أقرب مديقل

العروض الدعوية

المكتبة

الرئيسية

## القصائد ملتقى اداعيات

- ملتقى اداعيات
- د. أميمة الجلاهمة
- أفراح الحميضي
- رقية المحارب
- د. بنت الرسالة
- أمها الجريس
- فاطمة البطاح
- ندى اليوسفي
- خيرية الحارثي
- د. نهى قاطرجي
- فوزية الخليوي
- شيماء علي
- إكرام الأزدي
- نور الجندلي
- عطاء أم معاذ
- د. جواهر آل الشيخ
- صباح الضامن
- سحر لثان
- صاحبة قلم
- لثان شريف

## ملتقى الداعيات

### الدكتورة أميمة بنت أحمد الجلاهمة صفحة

### الدكتورة أفراح بنت علي الحميضي صفحة

### الدكتورة رقية بنت محمد المحارب صفحة

### الدكتورة سحر لثان صفحة

## القصائد الرئيسية

- اعرف نبيك
- العلماء وطلبة العلم
- أفكار دعوية
- مكتبة صيد الفوائد
- الأنشطة الدعوية
- زاد الداعية
- زاد الخطيب
- العروض الدعوية
- للنساء فقط
- ملتقى اداعيات
- رسائل دعوية
- الفلاشات - القصص
- مقالات - تفريعات
- واحة الأدب
- منوعات - مختارات
- المثل والنحل
- الطيب الدعوية
- بحوث علمية
- تربية الأبناء
- سعادة الشريعة



saaaid.net

# صيد الفوائد

صيد الفوائد



البحث

ساهم معنا

اتصل بنا

أقرب مدقق

العروض الدعوية

المكتبة

الرئيسية

صيد الفوائد أميمة الجلاهمة

- مقالات
- الصفحة الرئيسية
- منتقى الداعيات
- للنساء فقط

صيد الفوائد مواقع اسلامية



## الدكتورة صفحة أميمة بنت أحمد الجلاهمة



لنمكن شبابنا من تجسيد الانتماء بشكل إيجابي  
ما حال المركز الدولي لمكافحة الإرهاب  
ارحل غير مأسوف عليك  
تحالف حبر الأعداء والأصدقاء  
السياب في الديمقراطية الأميركية  
تباطؤ غير مفهوم  
احترم تحترم  
كل منا معني بمكافحة الإرهاب  
أنحن من ندعم الإرهاب  
نتنظر صوتا يماثل صوت الشيخ الحسيني  
إجراءات ضرورية لا اختيارية

صيد الفوائد الرئيسية

- اعرف نبيك
- العلماء وطبقة العلم
- أفكار دعوية
- مكتبة صيد الفوائد
- الأنشطة الدعوية
- زاد الداعية
- زاد الخطيب
- العروض الدعوية
- للنساء فقط
- منتقى الداعيات
- رسائل دعوية
- الفلاشات - القصص
- مقالات - تغريدات
- واحة الأدب
- منوعات - مختارات
- المثل والنحل
- الطيب الداعية
- بحوث علمية
- تربية الأبناء
- سيادة الشريعة



saaid.net

# صيد الفوائد

صيد الفوائد



البحث

ساهم معنا

اتصل بنا

أقرب مدقق

العروض الدعوية

المكتبة

الرئيسية

صيد الفوائد أميمة الجلاهية

- مقالات
- الصفحة الرئيسية
- منتقى الداعيات
- للنساء فقط

صيد الفوائد مواقع اسلامية



## الدكتورة أميمة بنت أحمد الجلاهية



صيد الفوائد الرئيسية

- اعرف نبيك
- العلماء وطبئة العلم
- أفكار دعوية
- مكتبة صيد الفوائد
- الأنشطة الدعوية
- زاد الداعية
- زاد الخطيب
- العروض الدعوية
- للنساء فقط
- منتقى الداعيات
- رسائل دعوية
- الفلاشات - القصص
- مقالات - تغريدات
- واحة الأدب
- منوعات - مختارات
- المثل والنحل
- الطيب الدعوية
- بحوث علمية
- تربية الأبناء
- سيادة الشريعة

لنمكن شبابنا من تجسيد الانتماء بشكل إيجابي  
ما حال المركز الدولي لمكافحة الإرهاب  
ارحل غير مأسوف عليك  
تحالف حير الأعداء والأصدقاء  
السياب في الديمقراطية الأميركية  
تباطؤ غير مفهوم  
احترم تحترم  
كل منا معني بمكافحة الإرهاب  
أنحن من ندعم الإرهاب  
نتنظر صوتا يماثل صوت الشيخ الحسيني  
إجراءات ضرورية لا اختيارية



TWEETS  
19.6K

FOLLOWING  
256

FOLLOWERS  
19.7K

LIKES  
5,454

أميمة أحمد الجلاهمة

@OmaimaAlJalahma

كاتبة و أسندة مشارك في جلمعة الادماء، نخصص

عقيدة، مقارنة الأدبن

Saudi Arabia

Joined October 2011

52 Photos and videos



Tweets

Tweets & replies

Media



Pinned Tweet



@OmaimaAlJalahma أميمة أحمد الجلاهمة · Jun 22

(رَبِّ اغْفِرْ لِي وَلِوَالِدَيَّ وَلِمَن دَخَلَ بَيْتِي مُؤْمِنًا وَلِلْمُؤْمِنِينَ  
وَالْمُؤْمِنَاتِ وَلَا تَزِدِ الظَّالِمِينَ إِلَّا تَبَارًا)



45



25



أميمة أحمد الجلاهمة Retweeted



@asuwayed عبدالعزيز السويد · 23h

دي مبسورا : ما بجري في مدينة #حلب برفى إلى جرائم حرب  
#روسيا #سوريا



10



أميمة أحمد الجلاهمة Retweeted



@Naifchair كرسى الأمير نايف · Sep 25

naifchair.kau.edu.sa/Pages-N02.aspx

أميمة أحمد الجلاهمة

@OmaimaAlJalahma

كاتبة و أساتذة مشارك في جامعة الدمام، تخصص  
عقيدة، مقارنة الأدب

Saudi Arabia

Joined October 2011

52 Photos and videos



تقبل الله منا  
ومنكم ..وعيدكم مبارك



Twee







saaaid.net

# صيد الفوائد

صيد الفوائد



البحث

ساهم معنا

اتصل بنا

أقرب مديقل

العروض الدعوية

المكتبة

الرئيسية

## القصائد ملتقى لداعيات

- ملتقى اداعيات
- د. أميمة الجلاهمة
- أفراح الحميضي
- رقية المحارب
- د. بنت الرسالة
- أمها الجريس
- فاطمة البطاح
- ندى اليوسفي
- خيرية الحارثي
- د. نهى قاطرجي
- فوزية الخليوي
- شيماء علي
- إكرام الأزدي
- نور الجندلي
- عطاء أم معاذ
- د. جواهر آل الشيخ
- صباح الضامن
- سحر لثان
- صاحبة قلم
- لثان شريف

## ملتقى الداعيات

## الدكتورة أميمة بنت أحمد الجلاهمة

## الدكتورة أفراح بنت علي الحميضي

## الدكتورة رقية بنت محمد المحارب

## الدكتورة

## القصائد الرئيسية

- اعرف نبيك
- العلماء وطلبة العلم
- أفكار دعوية
- مكتبة صيد الفوائد
- الأنشطة الدعوية
- زاد الداعية
- زاد الخطيب
- العروض الدعوية
- للنساء فقط
- ملتقى اداعيات
- رسائل دعوية
- الفلاشات - القصص
- مقالات - تفريعات
- واحة الأدب
- منوعات - مختارات
- المثل والنحل
- الطيب الدعوية
- بحوث علمية
- تربية الأبناء
- سعادة الشريعة



saaaid.net

# صيد الفوائد

صيد الفوائد



البحث

ساهم معنا

اتصل بنا

أقرب مديقل

العروض الدعوية

المكتبة

الرئيسية

## القصص المتنوعة لاداعات

- منتقى اداعات
- د. أميمة الجلاهية
- أفراح الحميضي
- رقية المحارب
- د. بنت الرسالة
- أمها الجريس
- فاطمة البطاح
- ندى اليوسفي
- خيرية الحارثي
- د. نهى قاطرجي
- فوزية الخليوي
- شيماء علي
- إكرام الزيد
- نور الجندلي
- عطاء أم معاذ
- د. جواهر آل الشيخ
- صباح الضامن
- سحر لبيان
- صاحبة قلم
- لينا شرف

## منتقى الداعيات

## الدكتورة أميمة بنت أحمد الجلاهية

## الدكتورة أفراح بنت علي الحميضي

## الدكتورة رقية بنت محمد المحارب

## الدكتورة

## القصص المتنوعة لاداعات

- اعرف نبيك
- العلماء وطبقة العلم
- أفكار دعوية
- مكتبة صيد الفوائد
- الأنشطة الدعوية
- زاد الداعية
- زاد الخطيب
- العروض الدعوية
- للنساء فقط
- منتقى اداعات
- رسائل دعوية
- الفلاشات - القصص
- مقالات - تفريعات
- واحة الأدب
- منوعات - مختارات
- المثل والنحل
- الطيب الدعوية
- بحث علمية
- تربية الأبناء
- سعادة الشريعة

## صفحة الدكتور افراح بنت علي الحميضي

### الفوائد الرئيسية

- اعرف نبيك
- العلماء وطبقة العلم
- أفكار دعوية
- مكتبة صيد الفوائد
- الأنشطة الدعوية
- زاد الداعية
- زاد الخطيب
- العروض الدعوية
- للنساء فقط
- منتقى الداعيات
- رسائل دعوية
- الفلاشات - القصص
- مقالات - تغريدات
- واحة الأدب
- منوعات - مختارات
- المثل والنحل
- الطبيب الداعية
- بحوث علمية
- تربية الأبناء
- سعادة الشريعة

أورام البناء الثقافي

بيدين، بكتفين، بخطوتين معاً

للنساء: جرعة فقه تقى من غوائل الجهل

هوس نسائي..!

لماذا .. 8 مارس؟؟

معلمة بشهادة جلاد!!

محصلة التشبّه: عرى الأجساد، خواء العقول!!

الطلاق ونعبة النّار

Power Rangers والبقية ستأتي..

أصوات تطرب لها الذناب!!

ما لم يدركه الرّقيب

### الفوائد فرح لميضي

- مقالات
- الصفحة الرئيسية
- منتقى الداعيات
- للنساء فقط

### الفوائد مواقع اسلامية

الآن

الموقع باللغة الإنجليزية

الأسهم  
والمعاملات المالية



saaaid.net

# صيد الفوائد

صيد الفوائد



البحث

ساهم معنا

اتصل بنا

أخبر صديقك

العروض الدعوية

المكتبة

الرئيسية

## صفحة الدكتوراه افراح بنت علي الحميضي

الرئيسية

الفوائد

- اعرف نبيك
- العلماء وطبقة العلم
- أفكار دعوية
- مكتبة صيد الفوائد
- الأنشطة الدعوية
- زاد الداعية
- زاد الخطيب
- العروض الدعوية
- للنساء فقط
- منتقى الداعيات
- رسائل دعوية
- الفلاشات - القصص
- مقالات - تغريدات
- واحة الأدب
- منوعات - مختارات
- المثل والنحل
- الطيب الدعوية
- بحوث علمية
- تربية الأبناء
- سعادة الشريعة

الفوائد فرح لحميضي

- مقالات
- الصفحة الرئيسية
- منتقى الداعيات
- للنساء فقط

الفوائد مواقع اسلامية

الآن

الموقع باللغة الإنجليزية

الأسهم  
والمعاملات المالية

أورام البناء الثقافي

بيدين، بكتفين، بخطوتين معاً

للنساء: جرعة فقه تقى من غوائل الجهل

هوس نسائي..!

لماذا .. 8 مارس؟؟

معلمة بشهادة جلاد!!

محصلة التشبّه: عرى الأجساد، خواء العقول!!

الطلاق ولعبة النّار

Power Rangers واليقظة ستأتي..

أصوات تطرب لها الذناب!!

ما لم يدركه الرّقيب

## Power Rangers والبقية ستأتي..

د. أفراح بنت علي الحميضي

(Power Rangers) فيلم أجنبي مُترجم يخدم فئة الطفولة المبكرة ومرحلة سن المراهقة (سن 10-20) يطرح العديد من الأفكار المتناقضة لمعالم الأديان، وما سأعرضه هنا كمثال للعديد من الأفلام التي توجد بين يدي الأطفال والمراهقين بعلم وغفلة من الوالدين أو بجهل منهم، وهو ينتشر في أسواقنا وتنتفقه أيدي النشء، ليصبح في نفوسهم - هو وأشبابه الكثير - نماذج سيئة وقيما مدمرة؛ أقلها تمجيد الإنسان الغربي ومنحه قوة تنافس، بل تتفوق على قوة الله العظيم - أسْتَغْفِرُ اللهَ - .

تتمحور قصة هذا الفيلم حول مجموعة من الفتيان والفتيات يصارعون قوى الشر (الشيطان) وهم يخوضون لأجل تحقيق النصر عليه وعلى أعوانه العديد من المعارك، تنتهي بانتصارهم في نهاية المعركة.

- اعرف نبيك
- العلماء وطبقة العلم
- أفكار دعوية
- مكتبة صيد الفوائد
- الأنشطة الدعوية
- زاد الداعية
- زاد الخطيب
- العروض الدعوية
- للنساء فقط
- منتقى الداعيات
- رسائل دعوية
- الفلاشات - القصص
- مقالات - تفريغات
- واحة الأدب
- منوعات - مختارات
- المثل والنحل
- الطيب الدعوية
- بحث علمية
- تربية الأبناء
- سبادة الشريعة

# How the Internet Helps Women Gain Authority in Islam

---

Why are 43 female preachers on a Salafi website?

Answer: Internet outreach trumps gender norms.

Evidence: **text** and **tweets**

- Saaid.net administrators solicited women.
  - **email interviews**
- Women use citations (the Salafi method) less than men.
  - **topic model, word frequencies in 21K docs**
- Women use identity authority more than men.
  - **reading (qual), personal pronoun use (quant)**
- Women reach new audiences online.
  - **retweets**

# Introduction

## Other scholars' research

- What gets censored in China?
- Do far right protests in Europe increase support for terrorism?
- Where are civilians being killed in Syria?
- How does fake news spread online?
- Did Shakespeare write all of Shakespeare's plays?
- Many, many more...

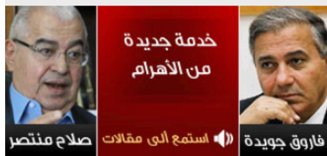
# Workshop Overview

We will work a single example in detail: Al-Ahram articles.



# Workshop Overview

- مؤتمر برعاية سوزان مبارك لدعم سيدات الأعمال
- فتح منفذ رفح 3 أيام ابتداء من 3 يناير المقبل
- نقابة جديدة للعاملين بالضرائب والمالية والجمارك
- زيارة نظيف لإثيوبيا مرحلة جديدة للعلاقات الثنائية
- هجانة جديدة بمدينة نصر.. أين القانون؟
- عودة 22 ألف لاجئ سوداني إلى بلادهم قريبا



المزيد

كتاب الأهرام

حالة طوارئ في إيران لمواجهة مظاهرات المعارضة  
أعلنت السلطات الإيرانية أمس حالة الطوارئ لمدة ثلاثة أيام، كما أعلنت قيادة الحرس الثوري في طهران حالة التأهب لمواجهة مظاهرات المعارضة.

[تعليقات - 1]

الأهرام يحتفل ببداية عامه الـ 135  
تحتفل مؤسسة الأهرام اليوم ببداية العام 135 على تأسيس شركة الأهرام في 27 ديسمبر سنة 1875. وبهذه المناسبة قرر الدكتور عبدالمنعم سعيد رئيس مجلس الإدارة الاحتفال في يوم 27 ديسمبر من كل عام بيوم الأهرام.

[تعليقات - 20]

موقع الأهرام في ثوب جديد من اليوم  
ابتداء من اليوم يطالع المترددون على موقع الأهرام الإلكتروني شكلا جديدا للموقع. في إطار المرحلة الأولى من مراحل تطويره المستمرة خلال الأشهر القليلة المقبلة.

[تعليقات - 279]

نظرة الثقافة

لمرأة والطفل

ذاكرة وتلفزيون

لكتاب

لأعمدة

رأى حرة

ملفات الأهرام

ريد الأهرام

برلمان الثورة

لأخيرة

أبواب اسبوعية

سياحة وسفر

ملفات دولية

تدريب وتعليم

طب وعلوم وبيئة

ملحق الجمعة

صور برلمانية

فكر ديني

هوامش حرة

أوراق دبلوماسية

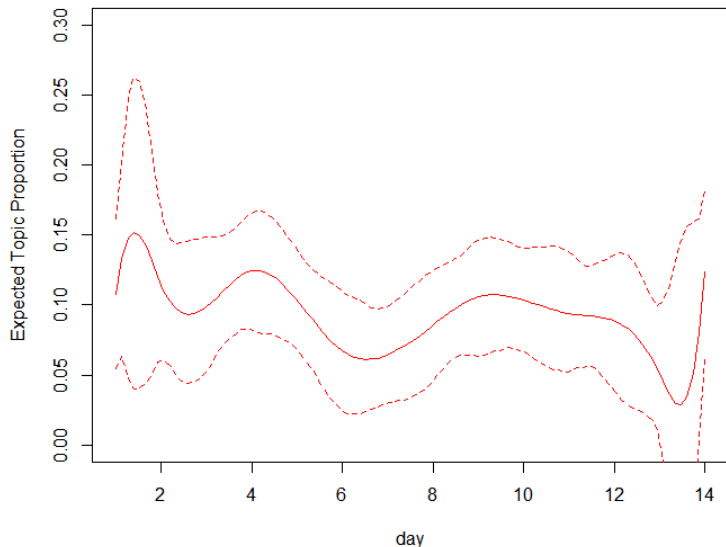
# Workshop Overview



# Workshop Overview

## Palestine

اسرائيل; حماس; الفلسطينية; مصالح; الحدود; اليهودية; الفلسطينين; المفاوضات; القدس; الحزب



# Workshop Overview

We will work a single example in detail: Al-Ahram articles.

1. Principles of text analysis research.
2. Acquiring text.
3. Pre-processesing text for quantitative analysis.
4. Topic Modeling (Unsupervised machine learning).
5. Classification (Supervised machine learning).
6. Visualization.

# Workshop Overview

We will work a single example in detail: Al-Ahram articles.

1. **Principles of text analysis research.**
2. Acquiring text.
3. Pre-processesing text for quantitative analysis.
4. Topic Modeling (Unsupervised machine learning).
5. Classification (Supervised machine learning).
6. Visualization.

# Principles of Text Analysis Research

## Research Design

- Puzzle/Question (a “Topic” is not enough)
- Theory and hypotheses
- Concepts
- Measurement
- Testing
- Report findings

# Principles of Text Analysis Research

What is quantitative text analysis good for?

- Concepts
- Measurement
- Testing

What would you do with infinite time?

Too often, text analysis starts with texts. Start with a question!

# Principles of Text Analysis Research

## Four Principles of Quantitative Text Analysis (Grimmer and Stewart, 2013)

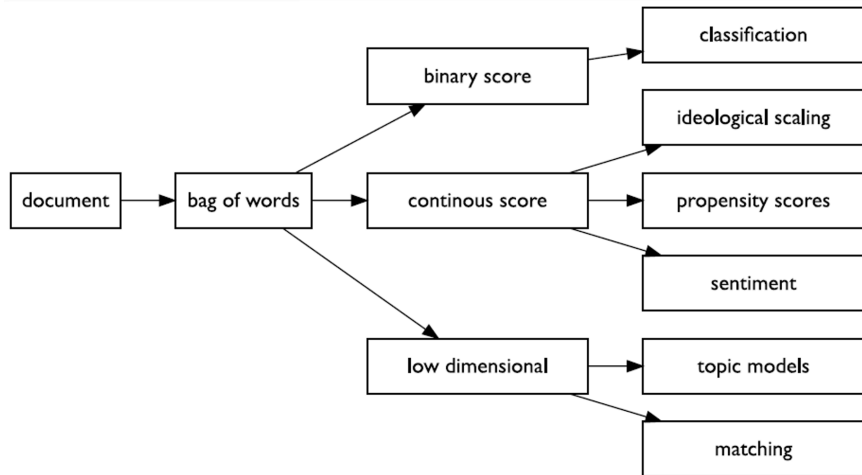
- 
- (1) All quantitative models of language are wrong—but some are useful.
  - (2) Quantitative methods for text amplify resources and augment humans.
  - (3) There is no globally best method for automated text analysis.
  - (4) Validate, Validate, Validate.
- 

### My additions:

- (5) All text analysis is dimension reduction.
- (6) There is no free lunch.  
(supervised methods require front-end work, unsupervised require back-end work)

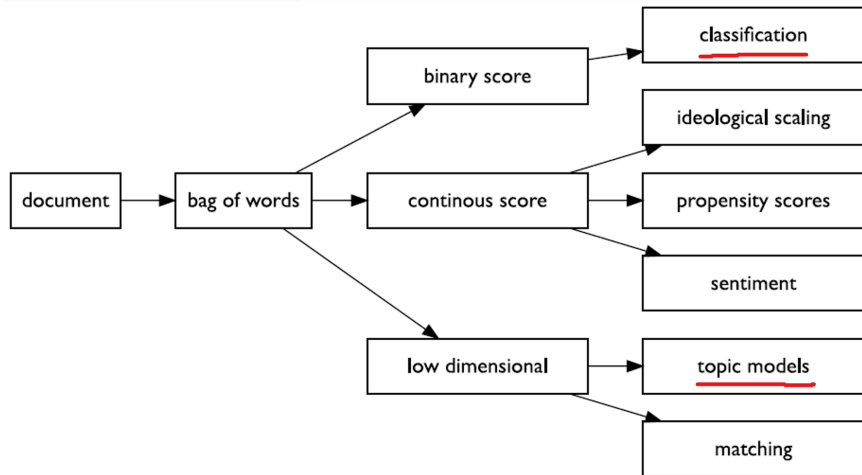


# Principles of Text Analysis Research



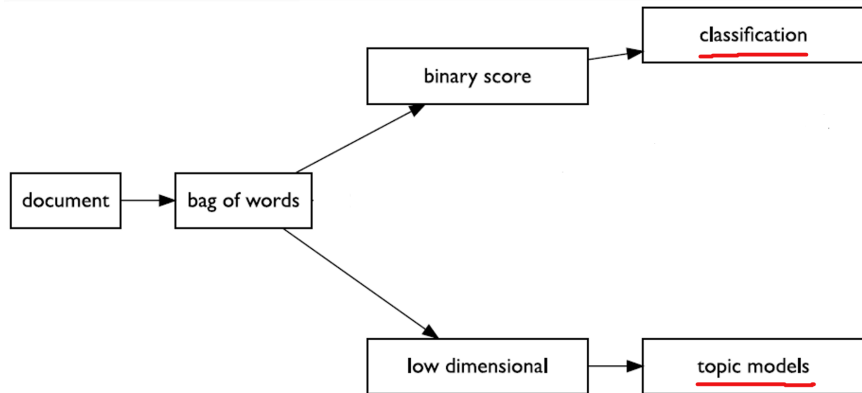
2nd generation text analysis

# Principles of Text Analysis Research



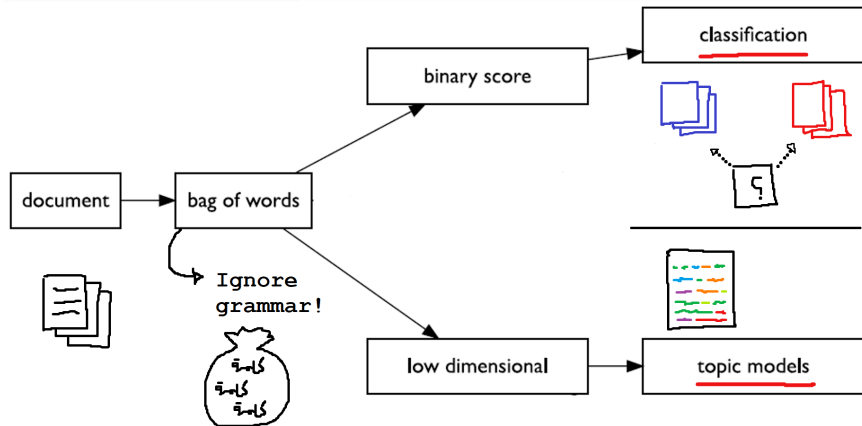
2nd generation text analysis

# Principles of Text Analysis Research



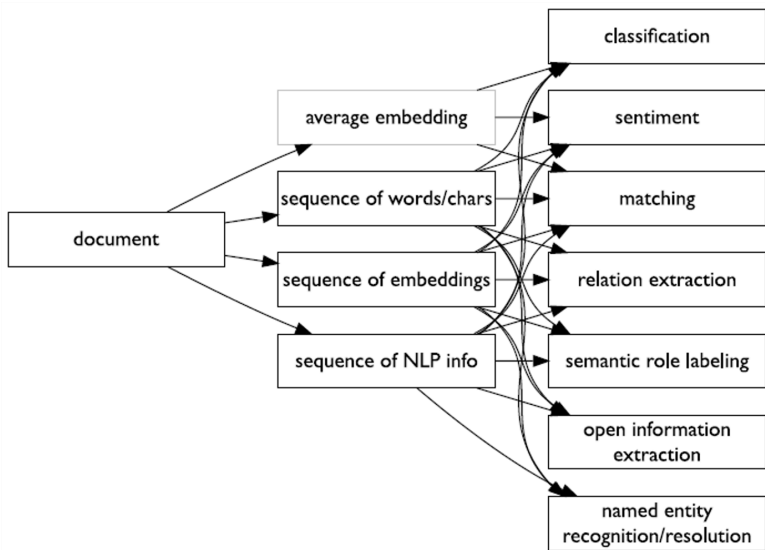
2nd generation text analysis

# Principles of Text Analysis Research



2nd generation text analysis

# Principles of Text Analysis Research



3rd generation text analysis

# Principles of Text Analysis Research

## Key conclusions

- Text analysis is dimension reduction.
- Speeds up repetitive coding tasks.
- Can help quantitative and qualitative research.
- Anyone who says it is magic is selling you something.

# Workshop Overview

1. Principles of text analysis research.
2. **Acquiring text.**
3. Pre-processesing text for quantitative analysis.
4. Topic Modeling (Unsupervised machine learning).
5. Classification (Supervised machine learning).
6. Visualization.

# Acquiring Text


The goal: .txt files in a directory



# Acquiring Text

The goal: .txt files in a directory

- مؤتمر برعاية سوزان مبارك لدعم سيدات الأعمال
- فتح منفذ رفح 3 أيام ابتداء من3 يناير المقبل
- نقابة جديدة للعاملين بالضرائب والمالية والجمارك
- زيارة نظيف لإثيوبيا مرحلة جديدة للعلاقات الثنائية
- هجانة جديدة بمدينة نصر.. أين القانون؟
- عودة22 ألف لاجئ سوداني إلى بلادهم قريبا



**خدمة جديدة  
من الأهرام**

صلاح منتصر



**فاروق جويده**

استمع الى مقالات

المزيد

كتاب الأهرام

## حالة طوارئ في إيران لمواجهة مظاهرات المعارضة

أعلنت السلطات الإيرانية أمس حالة الطوارئ لمدة ثلاثة أيام، كما أعلنت قيادة الحرس الثوري في طهران حالة التأهب لمواجهة مظاهرات المعارضة.

[تعليقات - 1]

الأهرام يحتفل ببدء عامه الـ 135  
تحتفل مؤسسة الأهرام اليوم ببدء العام 135 على تأسيس شركة الأهرام في 27 ديسمبر سنة 1875. وبهذه المناسبة قرر الدكتور عبدالمنعم سعيد رئيس مجلس الإدارة الاحتفال في يوم 27 ديسمبر من كل عام بيوم الأهرام.

[تعليقات - 20]

موقع الأهرام في ثوب جديد من اليوم  
ابتداء من اليوم يطالع المترددون على موقع الأهرام الإلكتروني شكلا جديدا للموقع. في إطار المرحلة الأولى من مراحل تطويره المستمرة خلال الأشهر القليلة المقبلة.

[تعليقات - 279]

دنيا الثقافة

المرأة والطفل

إذاعة وتلفزيون

الكتاب

الإعلام

أراء حرة

ملفات الأهرام

بريد الأهرام

برلمان الثورة

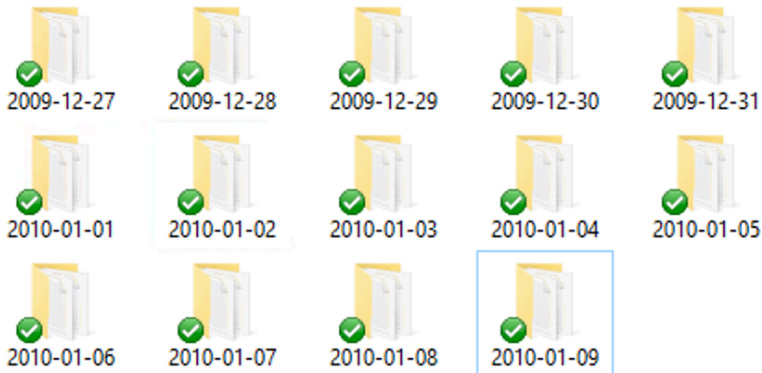
الاخيرة

ابواب اسبوعية

- سباحة وسفر
- ملفات دولية
- تنديب وتعليم
- طب وعلوم وثيقة
- ملحق الجمعة
- صور برلمانية
- فكر ديني
- هوامش حرة
- أوراق دبلوماسية










# Acquiring Text

The goal: .txt files in a directory



## Acquiring Text

The goal: .txt files in a directory

-  article\_2\_1106.txt
-  article\_2\_1107.txt
-  article\_2\_1112.txt
-  article\_2\_1114.txt
-  article\_3\_1227.txt
-  article\_3\_1228.txt
-  article\_3\_1236.txt
-  article\_4\_1147.txt
-  article\_4\_1150.txt

# Acquiring Text

The goal: .txt files in a directory

Three types of text sources:

1. Digital
2. Online Analog
3. Offline Analog

# Acquiring Text

Steps:

- Identify text source
- Explore the structure
- Navigate the structure
- Get the text onto your computer
- Clean the text

# Acquiring Text

Steps: Digital

- Identify text source:  
website, digital files
- Explore the structure:  
hyperlinks, html
- Navigate the structure:  
regex, spider
- Get the text onto your computer:  
scraping, API
- Clean the text:  
regex, html parser

# Acquiring Text

Steps: Digital, Online Analog

- Identify text source:  
website, digital files, digital files
- Explore the structure:  
hyperlinks, html, hyperlinks, html
- Navigate the structure:  
regex, spider, regex, spider
- Get the text onto your computer:  
scraping, API, scraping, OCR
- Clean the text:  
regex, html parser, OCR errors, doc format

# Acquiring Text

Steps: Digital, Online Analog, Offline Analog

- Identify text source:  
website, digital files, digital files, library, archive
- Explore the structure:  
hyperlinks, html, hyperlinks, html, by hand
- Navigate the structure:  
regex, spider, regex, spider, by hand
- Get the text onto your computer:  
scraping, API, scraping, OCR, photography, OCR
- Clean the text:  
regex, html parser, OCR errors, doc format, OCR errors,  
doc format



# Acquiring Text: Installing R

---

<https://www.r-project.org/>



[\[Home\]](#)

## Download

[CRAN](#)

## R Project

[About R](#)

[Logo](#)

[Contributors](#)

[What's New?](#)

[Reporting Bugs](#)

[Development Site](#)

[Conferences](#)

[Search](#)

## R Foundation

[Foundation](#)

[Board](#)

[Members](#)

[Donors](#)

[Donate](#)

## The R Project for Statistical Computing

### Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

### News

- **R version 3.4.3 (Kite-Eating Tree) prerelease versions** will appear starting Monday 2017-11-20. Final release is scheduled for Thursday 2017-11-30.
- **R version 3.4.2 (Short Summer)** has been released on Thursday 2017-09-28.
- **The R Journal Volume 9/1** is available.
- **R version 3.3.3 (Another Canoe)** has been released on Monday 2017-03-06.
- **The R Journal Volume 8/2** is available.
- **useR! 2017** (July 4 - 7 in Brussels) has opened registration and more at <http://user2017.brussels/>
- Tomas Kalibera has joined the R core team.
- The R Foundation welcomes five new ordinary members: Jennifer Bryan, Dianne Cook, Julie Josse,

# Acquiring Text: Installing R

---

<https://www.r-project.org/>

## CRAN Mirrors

The Comprehensive R Archive Network is available at the following URLs, please choose a location close to you. Some statistics on the status of the [windows release](#), [windows old release](#).

If you want to host a new mirror at your institution, please have a look at the [CRAN Mirror HOWTO](#).

### 0-Cloud

<https://cloud.r-project.org/>

<http://cloud.r-project.org/>



Automatic redirection to servers worldwide, currently sponsored

Automatic redirection to servers worldwide, currently sponsored

### Algeria

<https://cran.usthb.dz/>

<http://cran.usthb.dz/>

University of Science and Technology Houari Boumediene

University of Science and Technology Houari Boumediene

### Argentina

<http://mirror.fcaglp.unlp.edu.ar/CRAN/>

Universidad Nacional de La Plata

### Australia

<https://cran.csiro.au/>

<http://cran.csiro.au/>

<https://mirror.aarnet.edu.au/pub/CRAN/>

<https://cran.ms.unimelb.edu.au/>

<https://cran.curtin.edu.au/>

CSIRO

CSIRO

AARNET

School of Mathematics and Statistics, University of Melbourne

Curtin University of Technology

### Austria

<https://cran.wu.ac.at/>

<http://cran.wu.ac.at/>

Wirtschaftsuniversität Wien

Wirtschaftsuniversität Wien

### Belgium

<http://www.freeststatistics.org/cran/>

<https://lib.ugent.be/CRAN/>

<http://lib.ugent.be/CRAN/>

K.U.Leuven Association

Ghent University Library

Ghent University Library

# Acquiring Text: Installing R

---

<https://www.r-project.org/>



CRAN

[Mirrors](#)

[What's new?](#)

[Task Views](#)

[Search](#)

About R

[R Homepage](#)

[The R Journal](#)

Software

[R Sources](#)

[R Binaries](#)

[Packages](#)

[Other](#)

Documentation

[Manuals](#)

[FAQs](#)

[Contributed](#)

## The Comprehensive R Archive Network

### Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely need one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

### Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably want to do it!

- The latest release (Thursday 2017-09-28, Short Summer) [R-3.4.2.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features](#) and [bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

### Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are,

# Acquiring Text: Installing R

---

<https://www.r-project.org/>



CRAN

[Mirrors](#)

[What's new?](#)

[Task Views](#)

[Search](#)

About R

[R Homepage](#)

[The R Journal](#)

Software

[R Sources](#)

[R Binaries](#)

[Packages](#)

[Other](#)

Documentation

[Manuals](#)

[FAQs](#)

[Contributed](#)

## R for Windows

Subdirectories:

[base](#)

[contrib](#)

[old contrib](#)

[Rtools](#)

Binaries for base distribution. This is what you want to [install R for the first time](#).

Binaries of contributed CRAN packages (for R  $\geq$  2.13.x; managed by Uwe Ligges). There is also information on [third party software](#) available for CRAN Windows service and corresponding environment and make variables.

Binaries of contributed CRAN packages for outdated versions of R (for R  $<$  2.13.x; managed by Uwe Ligges).

Tools to build R and R packages. This is what you want to build your own packages on Windows, or to build itself.

Please do not submit binaries to CRAN. Package developers might want to contact Uwe Ligges directly in case of questions / suggestions for Windows binaries.

You may also want to read the [R FAQ](#) and [R for Windows FAQ](#).

Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded

# Acquiring Text: Installing R

---

<https://www.r-project.org/>



CRAN

[Mirrors](#)  
[What's new?](#)  
[Task Views](#)  
[Search](#)

About R

[R Homepage](#)  
[The R Journal](#)

Software

[R Sources](#)  
[R Binaries](#)  
[Packages](#)  
[Other](#)

Documentation

[Manuals](#)  
[FAQs](#)  
[Contributed](#)

R-3.4.2 for Windows (32/64 bit)

[Download R 3.4.2 for Windows](#) 75 megabytes, 32/64 bit)

[Installation and other instructions](#)  
[New features in this version](#)

If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the [md5sum](#) the [fingerprint](#) on the master server. You will need a version of md5sum for windows: both [graphical](#) and [command line versions](#) are available.

Frequently asked questions

- [Does R run under my version of Windows?](#)
- [How do I update packages in my previous version of R?](#)
- [Should I run 32-bit or 64-bit R?](#)

Please see the [R FAQ](#) for general information about R and the [R Windows FAQ](#) for Windows-specific information.

Other builds

- Patches to this release are incorporated in the [r-patched snapshot build](#).
- A build of the development version (which will eventually become the next major release of R) is available in the [r-devel snapshot build](#)
- [Previous releases](#)

Note to webmasters: A stable link which will redirect to the current Windows binary release is [<CRAN MIRROR>/bin/windows/base/release.htm](https://cran.r-project.org/bin/windows/base/release.htm).

---

Last change: 2017-09-28, by Duncan Murdoch

# Acquiring Text: Installing R

<https://www.r-project.org/>



CRAN  
[Mirrors](#)  
[What's new?](#)  
[Task Views](#)  
[Search](#)

About R  
[R Homepage](#)  
[The R Journal](#)

Software  
[R Sources](#)  
[R Binaries](#)  
[Packages](#)  
[Other](#)

Documentation  
[Manuals](#)  
[FAQs](#)  
[Contributed](#)

## R for Mac OS X

This directory contains binaries for a base distribution and packages to run on Mac OS X (release 10.6 and above). Mac OS 8.6 to 9.2 (and Mac OS 9.2 to 9.3) are no longer supported but you can find the last supported release of R for these systems (which is R 1.7.1) [here](#). Releases for old Mac OS X systems (through Mac OS X 10.5) and PowerPC Macs can be found in the [old](#) directory.

Note: CRAN does not have Mac OS X systems and cannot check these binaries for viruses. Although we take precautions when assembling binaries, you must use the normal precautions with downloaded executables.

As of 2016/03/01 package binaries for R versions older than 2.12.0 are only available from the [CRAN archive](#) so users of such versions should set their CRAN mirror setting accordingly.

### R 3.4.2 "Short Summer" released on 2017/09/28

**Important:** since R 3.4.0 release we are now providing binaries for OS X 10.11 (El Capitan) and higher using non-Apple toolkit to provide support for OpenMP and C++17 standard features. Please read the corresponding note below.

Please check the MD5 checksum of the downloaded image to ensure that it has not been tampered with or corrupted during the mirroring process.

example type

md5 R-3.4.2.pkg

in the *Terminal* application to print the MD5 checksum for the R-3.4.2.pkg image. On Mac OS X 10.7 and later you can also validate the signature

pkgutil --check-signature R-3.4.2.pkg

### Files:

**R 3.4.2** binary for OS X 10.11 (El Capitan) and higher, signed package. Contains R 3.4.2 framework, R.app GUI 1.70 for Intel Macs, Tcl/Tk 8.6.6 X11 libraries and Texinfo 5.2. The latter two components are optional and can be omitted when choosing "custom install", they are only needed if you want to use the `tcltk` R package or build package documentation from sources.

Note: the use of X11 (including `tcltk`) requires [XQuartz](#) to be installed since it is no longer part of OS X. Always re-install XQuartz when upgrading your OS X to a new major version.

**Important:** this release uses Clang 4.0.0 and GNU Fortran 6.1, neither of which is supplied by Apple. If you wish to compile R packages from sources, you will need to download and install

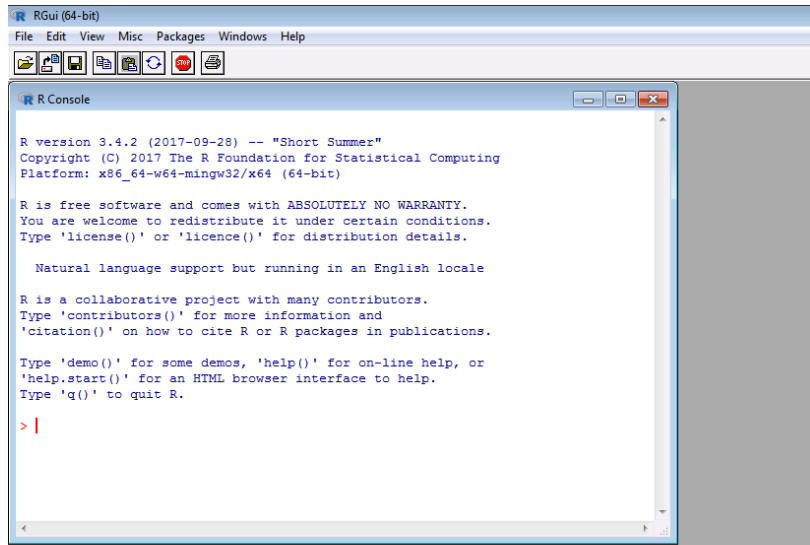
### R-3.4.2.pkg

MD5-hash: 71d2db804c38cf2ce8f9cfed1fa860  
SHA-hash: c577a8fc006e5a230b2932545340e9e1273dha  
(ca. 61MB)

# Acquiring Text: Installing R

---

## R console: Windows



The screenshot shows the RGui (64-bit) application window. The title bar reads "RGui (64-bit)". The menu bar includes "File", "Edit", "View", "Misc", "Packages", "Windows", and "Help". Below the menu bar is a toolbar with icons for file operations and execution. The "R Console" window is open, displaying the following text:

```
R version 3.4.2 (2017-09-28) -- "Short Summer"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

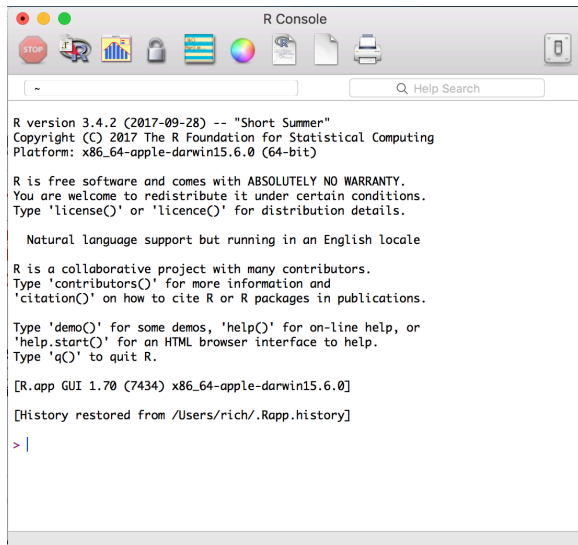
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

# Acquiring Text: Installing R

---

## R console: Mac



```
R version 3.4.2 (2017-09-28) -- "Short Summer"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

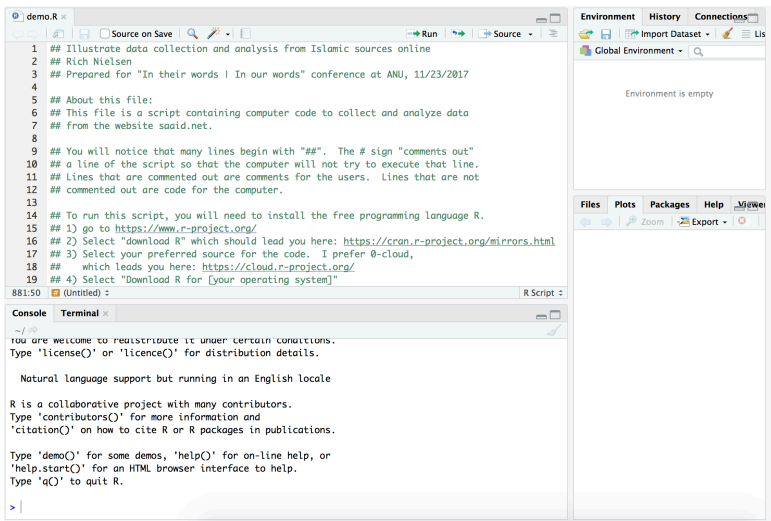
[R.app GUI 1.70 (7434) x86_64-apple-darwin15.6.0]
[History restored from /Users/rich/.Rapp.history]

> |
```



# Acquiring Text: Installing R

I highly recommend RStudio (I use it for Mac)  
(<https://www.rstudio.com/products/rstudio/download/>)



# Acquiring Text

R exercise.

- Hello world
- Objects and data structures
- `readLines()`
- regular expressions `grep()` and `gsub()`
- OCR with `tesseract`

Code: <http://www.mit.edu/~rnielsen/arabicTextWorkshop.zip>

Slides: [http://www.mit.edu/~rnielsen/arabic\\_text\\_slides.pdf](http://www.mit.edu/~rnielsen/arabic_text_slides.pdf)

<http://www.mit.edu/~rnielsen/> > "Helpful Stuff" (at the bottom)

# Workshop Overview

1. Principles of text analysis research.
2. Acquiring text.
3. **Pre-processesing text for quantitative analysis.**
4. Topic Modeling (Unsupervised machine learning).
5. Classification (Supervised machine learning).
6. Visualization.

# Turning Text into Data

Goal: Turn text into a **Document Term Matrix**.



## Turning Text into Data

```
> dtml[1:10,210:220]
```

	Karj	Klal	Kms	QT3	Qal	Qbl	Qdym	Qlb	Qlyl	SGyr	Sf7
1	0	0	0	0	0	0	0	0	0	0	0
2	1	3	4	1	3	1	1	1	1	1	1
3	0	0	0	0	0	0	1	0	0	0	0
4	0	0	1	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	1	2	0	1	0
7	0	0	0	0	0	0	0	0	0	0	0
8	0	0	3	0	3	0	0	0	0	0	7
9	2	1	0	0	2	0	0	0	0	0	3
10	0	0	0	0	1	0	3	0	2	1	3

# Turning Text into Data

Organizing files to make your life easier.

# Turning Text into Data

Organizing files to make your life easier.

Encodings and Unicode.



## Turning Text into Data

1. عندي كتاب ومقالة .
2. هذه المقالات جيدة .

# Turning Text into Data

1. عندي كتاب ومقالة .
2. هذه المقالات جيدة .

1. |عندي| |كتاب| |ومقالة| .
2. |هذه| |المقالات| |جيدة| .

Tokenize

# Turning Text into Data

1. عندي كتاب ومقالة .
2. هذه المقالات جيدة .

1. |عندي| |كتاب| |ومقالة| .
2. |هذه| |المقالات| |جيدة| .

Tokenize

Remove stopwords

# Turning Text into Data

1. عندي كتاب ومقالة .
2. هذه المقالات جيدة .

1. |عندي| |كتاب| |ومقالة| .
2. |هذه| |المقالات| |جيدة| .

Tokenize

Remove stopwords

Stemming

# Turning Text into Data

1. عندي كتاب ومقالة.  
2. هذه المقالات جيدة.

1. | عندي | كتاب | ومقالة | .  
2. | هذه | المقالات | جيدة | .

Tokenize  
Remove stopwords  
Stemming

	عند	كتاب	مقال	جيد
1.				
2.				

# Turning Text into Data

1. عندي كتاب ومقالة.  
2. هذه المقالات جيدة.

1. | عندي | كتاب | ومقالة | .  
2. | هذه | المقالات | جيدة | .

Tokenize

Remove stopwords

Stemming

	عند	كتاب	مقال	جيد
1.	1	1	1	0
2.	0	0	1	1

# Turning Text into Data

You can do a lot with a DTM, even without models.

- Word counts
- Visualizations
- Document similarity



saaaid.net

# صيد الفوائد

صيد الفوائد



البحث

ساهم معنا

اتصل بنا

أقرب مديقل

العروض الدعوية

المكتبة

الرئيسية

القصص  
القصص

- ملتقى انداعيات
- د. أميمة الجلاهمة
- أفراح الحميضي
- رقية المحارب
- د. بنت الرسالة
- أمها الجريس
- فاطمة البطاح
- ندى اليوسفي
- خيرية الحارثي
- د. نهى قاطرجي
- فوزية الخليوي
- شيماء علي
- إكرام الأزدي
- نور الجندي
- عطاء أم معاذ
- د. جواهر آل الشيخ
- صباح الضامن
- سحر لثان
- صاحبة قلم
- لثان شريف

ملتقى الداعيات

الدكتورة  
صفحة  
أميمة بنت أحمد الجلاهمة

الدكتورة  
صفحة  
أفراح بنت علي الحميضي

الدكتورة  
صفحة  
رقية بنت محمد المحارب

الدكتورة  
صفحة  
د. أميمة الجلاهمة

القصص  
الرئيسية

- اعرف نبيك
- العلماء وطلبة العلم
- أفكار دعوية
- مكتبة صيد الفوائد
- الأنشطة الدعوية
- زاد الداعية
- زاد الخطيب
- العروض الدعوية
- للنساء فقط
- ملتقى انداعيات
- رسائل دعوية
- الفلاشات - القصص
- مقالات - تفريعات
- واحة الأدب
- منوعات - مختارات
- المثل والنحل
- الطيب الدعوية
- بحوث علمية
- تربية الأبناء
- سعادة الشريعة



Term	% of documents using term		% of all words	
	Men	Women	Men	Women
<i>allāh</i>	95%	85%	3.0%	1.6%
<i>ṣalā allāh</i>	67	32	0.50	0.21
<i>raḍī allāh</i>	50	20	0.18	0.10
<i>ḥadīth</i>	64	39	0.30	0.14
<i>qāl/yaqūl (al-)rasūl</i>	30	10	0.047	0.026
Ibn Taymiyya	23	3.4	0.030	0.006
al-Bukhari	41	10	0.098	0.030
al-Bayhaqi	17	0.72	0.013	0.001
Abu Hurayra	29	3.6	0.048	0.008
al-Tabarani	16	1.2	0.014	0.002
Abu Dawud	29	3.6	0.043	0.006

**Table 2:** *Frequency of hadith-related phrases in men's and women's writing*

# Turning Text into Data

R Exercise. Code: <http://www.mit.edu/~rnielsen/arabicTextWorkshop.zip>

Slides: [http://www.mit.edu/~rnielsen/arabic\\_text\\_slides.pdf](http://www.mit.edu/~rnielsen/arabic_text_slides.pdf)  
<http://www.mit.edu/~rnielsen/> > “Helpful Stuff” (at the bottom)

# Workshop Overview

1. Principles of text analysis research.
2. Acquiring text.
3. Pre-processesing text for quantitative analysis.
4. **Topic Modeling (Unsupervised machine learning).**
5. Classification (Supervised machine learning).
6. Visualization.

# Topic Models

# Topic Models

Goal: Summarize a Document Term Matrix with Topics.

# Topic Models

Goal: Summarize a Document Term Matrix with Topics.

Good: low effort, quick

Bad: Not always sure what you get

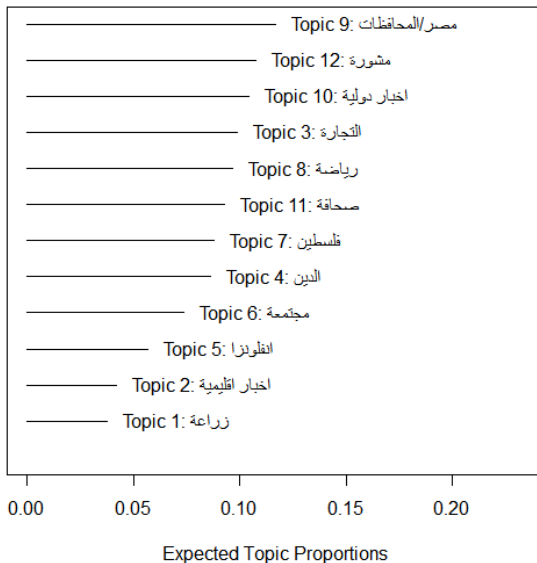
# Topic Models

```
> dtm1[1:10,210:220]
```

	Karj	Klal	Kms	QT3	Qal	Qbl	Qdym	Qlb	Qlyl	SGyr	Sf7
1	0	0	0	0	0	0	0	0	0	0	0
2	1	3	4	1	3	1	1	1	1	1	1
3	0	0	0	0	0	0	1	0	0	0	0
4	0	0	1	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	1	2	0	1	0
7	0	0	0	0	0	0	0	0	0	0	0
8	0	0	3	0	3	0	0	0	0	0	7
9	2	1	0	0	2	0	0	0	0	0	3
10	0	0	0	0	1	0	3	0	2	1	3

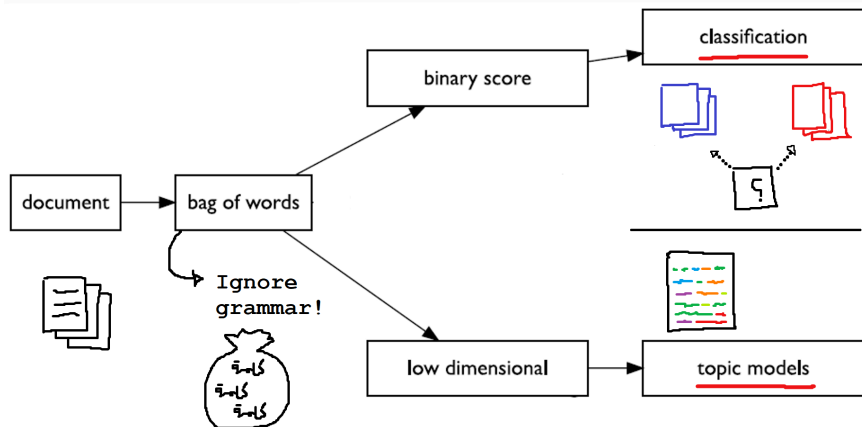
# Topic Models

## Top Topics





# Topic Models



# Topic Models

## Topics

gene 0.04  
dna 0.02  
genetic 0.01  
...

life 0.02  
evolve 0.01  
organism 0.01  
...

brain 0.04  
neuron 0.02  
nerve 0.01  
...

data 0.02  
number 0.02  
computer 0.01  
...

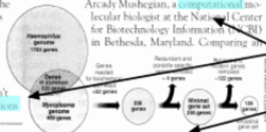
## Documents

### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Sir Anderson at the University in Sydney, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **numbers** game, particularly as more and more **genomes** are rapidly mapped and sequenced. "It may be a way of organizing any newly **sequenced genomes**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all

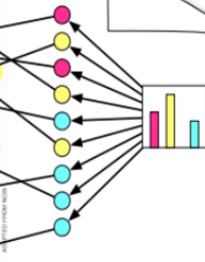


\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

## Topic proportions and assignments



# Topic Models

1. Pick the number of topics
2. Estimate the model
3. Interpret the topics
4. Summarize results

# Topic models

Two matrices estimated:

1) Topical Prevalence Matrix ( $D \times K$ )

2) Topical Content Matrix ( $V \times K$ )

# Topic models

Two matrices estimated:

1) Topical Prevalence Matrix ( $D \times K$ )

$$\theta = \begin{bmatrix} & \textit{Topic1} & \textit{Topic2} & \dots & \textit{TopicK} \\ \textit{Doc1} & .2 & .1 & \dots & 0.05 \\ \textit{Doc2} & .2 & .1 & \dots & .3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \textit{DocD} & 0 & 0 & \dots & .5 \end{bmatrix}$$

2) Topical Content Matrix ( $V \times K$ )

# Topic models

Two matrices estimated:

1) Topical Prevalence Matrix ( $D \times K$ )

$$\theta = \begin{bmatrix} & \textit{Topic1} & \textit{Topic2} & \dots & \textit{TopicK} \\ \textit{Doc1} & .2 & .1 & \dots & 0.05 \\ \textit{Doc2} & .2 & .1 & \dots & .3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \textit{DocD} & 0 & 0 & \dots & .5 \end{bmatrix}$$

2) Topical Content Matrix ( $V \times K$ )

$$\beta^T = \begin{bmatrix} & \textit{Topic1} & \textit{Topic2} & \dots & \textit{TopicK} \\ \textit{"text"} & .02 & .001 & \dots & 0.001 \\ \textit{"data"} & .001 & .02 & \dots & 0.001 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \textit{"analysis"} & .01 & .01 & \dots & 0.0005 \end{bmatrix}$$

# Topic models

Two matrices estimated:

$$X \approx \theta \beta$$

1) Topical Prevalence Matrix ( $D \times K$ )

$$\theta = \begin{bmatrix} & \textit{Topic1} & \textit{Topic2} & \dots & \textit{TopicK} \\ \textit{Doc1} & .2 & .1 & \dots & 0.05 \\ \textit{Doc2} & .2 & .1 & \dots & .3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \textit{DocD} & 0 & 0 & \dots & .5 \end{bmatrix}$$

2) Topical Content Matrix ( $V \times K$ )

$$\beta^T = \begin{bmatrix} & \textit{Topic1} & \textit{Topic2} & \dots & \textit{TopicK} \\ \textit{"text"} & .02 & .001 & \dots & 0.001 \\ \textit{"data"} & .001 & .02 & \dots & 0.001 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \textit{"analysis"} & .01 & .01 & \dots & 0.0005 \end{bmatrix}$$

# Topic models

For Latent Dirichlet Allocation...

- Consider document  $i$ , ( $i = 1, 2, \dots, D$ ).



# Topic models

For Latent Dirichlet Allocation...

- Consider document  $i$ , ( $i = 1, 2, \dots, D$ ).
- Suppose there are  $M_i$  total words and  $\mathbf{w}_i$  is an  $M_i \times 1$  vector, where  $w_{im}$  describes the  $m^{\text{th}}$  word used in the document.

# Topic models

For Latent Dirichlet Allocation...

- Consider document  $i$ , ( $i = 1, 2, \dots, D$ ).
- Suppose there are  $M_i$  total words and  $\mathbf{w}_i$  is an  $M_i \times 1$  vector, where  $w_{im}$  describes the  $m^{\text{th}}$  word used in the document.

$$\boldsymbol{\theta}_i | \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

# Topic models

For Latent Dirichlet Allocation...

- Consider document  $i$ , ( $i = 1, 2, \dots, D$ ).
- Suppose there are  $M_i$  total words and  $\mathbf{w}_i$  is an  $M_i \times 1$  vector, where  $w_{im}$  describes the  $m^{\text{th}}$  word used in the document.

$$\boldsymbol{\theta}_i | \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

$$\mathbf{z}_{im} | \boldsymbol{\theta}_i \sim \text{Multinomial}(1, \boldsymbol{\theta}_i)$$

# Topic models

For Latent Dirichlet Allocation...

- Consider document  $i$ , ( $i = 1, 2, \dots, D$ ).
- Suppose there are  $M_i$  total words and  $\mathbf{w}_i$  is an  $M_i \times 1$  vector, where  $w_{im}$  describes the  $m^{\text{th}}$  word used in the document.

$$\boldsymbol{\theta}_i | \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

$$\mathbf{z}_{im} | \boldsymbol{\theta}_i \sim \text{Multinomial}(1, \boldsymbol{\theta}_i)$$

$$w_{im} | \boldsymbol{\beta}_k, z_{imk} = 1 \sim \text{Multinomial}(1, \boldsymbol{\beta}_k)$$

# Topic models

For Latent Dirichlet Allocation...

- Consider document  $i$ , ( $i = 1, 2, \dots, D$ ).
- Suppose there are  $M_i$  total words and  $\mathbf{w}_i$  is an  $M_i \times 1$  vector, where  $w_{im}$  describes the  $m^{\text{th}}$  word used in the document.

$$\beta_k \sim \text{Dirichlet}(\mathbf{1})$$

$$\theta_i | \alpha \sim \text{Dirichlet}(\alpha)$$

$$\mathbf{z}_{im} | \theta_i \sim \text{Multinomial}(1, \theta_i)$$

$$w_{im} | \beta_k, z_{imk} = 1 \sim \text{Multinomial}(1, \beta_k)$$

# Topic models

For Latent Dirichlet Allocation...

- Consider document  $i$ , ( $i = 1, 2, \dots, D$ ).
- Suppose there are  $M_i$  total words and  $\mathbf{w}_i$  is an  $M_i \times 1$  vector, where  $w_{im}$  describes the  $m^{\text{th}}$  word used in the document.

$$\beta_k \sim \text{Dirichlet}(\mathbf{1})$$

$$\alpha_k \sim \text{Gamma}(\alpha, \beta)$$

$$\theta_i | \alpha \sim \text{Dirichlet}(\alpha)$$

$$\mathbf{z}_{im} | \theta_i \sim \text{Multinomial}(1, \theta_i)$$

$$w_{im} | \beta_k, z_{imk} = 1 \sim \text{Multinomial}(1, \beta_k)$$

# Topic models

For Latent Dirichlet Allocation...

- Consider document  $i$ , ( $i = 1, 2, \dots, D$ ).
- Suppose there are  $M_i$  total words and  $\mathbf{w}_i$  is an  $M_i \times 1$  vector, where  $w_{im}$  describes the  $m^{\text{th}}$  word used in the document.

$$\beta_k \sim \text{Dirichlet}(\mathbf{1})$$

$$\alpha_k \sim \text{Gamma}(\alpha, \beta)$$

$$\theta_i | \alpha \sim \text{Dirichlet}(\alpha)$$

$$\mathbf{z}_{im} | \theta_i \sim \text{Multinomial}(1, \theta_i)$$

$$w_{im} | \beta_k, z_{imk} = 1 \sim \text{Multinomial}(1, \beta_k)$$

Optimize with Variational Inference or Gibbs Sampling.

# Topic Models

R Exercise. Code: <http://www.mit.edu/~rnielsen/arabicTextWorkshop.zip>

Slides: [http://www.mit.edu/~rnielsen/arabic\\_text\\_slides.pdf](http://www.mit.edu/~rnielsen/arabic_text_slides.pdf)  
<http://www.mit.edu/~rnielsen/> > “Helpful Stuff” (at the bottom)



# Workshop Overview

1. Principles of text analysis research.
2. Acquiring text.
3. Pre-processesing text for quantitative analysis.
4. Topic Modeling (Unsupervised machine learning).
5. **Classification (Supervised machine learning).**
6. Visualization.

# Classification

Goal: Predict document type from text.

# Classification

Goal: Predict document type from text.

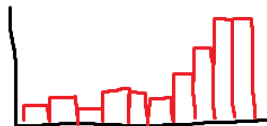
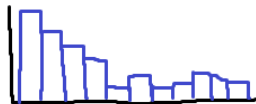
Good: You know what you are getting. Faster than hand-coding everything.

Bad: Time consuming to label documents.

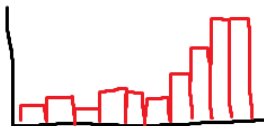
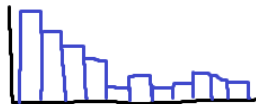
# Classification



# Classification



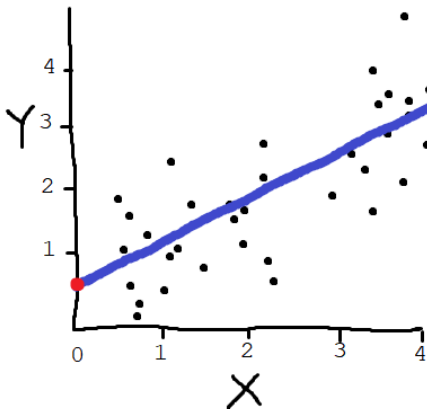
# Classification



# Classification

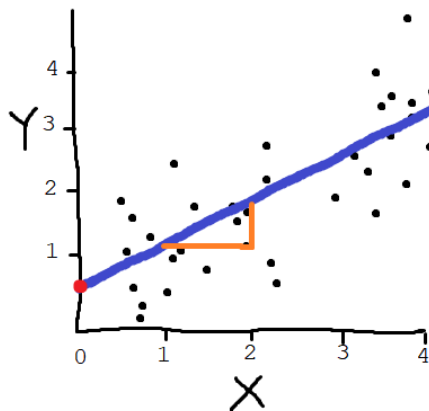


# Classification



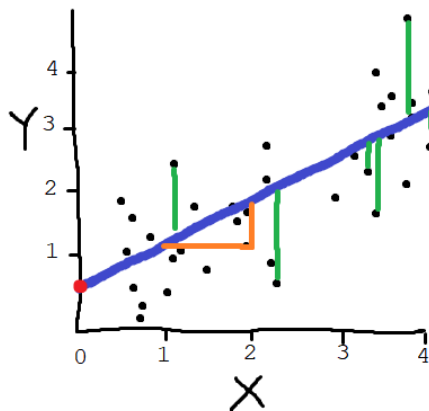


# Classification



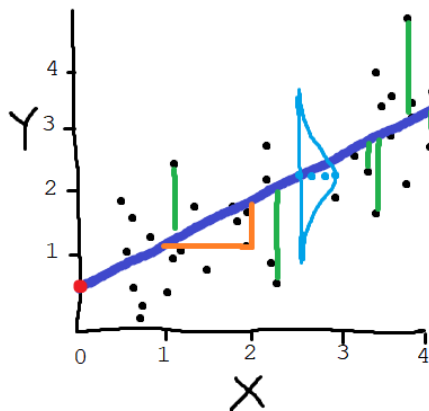
$$y = \alpha + \beta x$$

# Classification



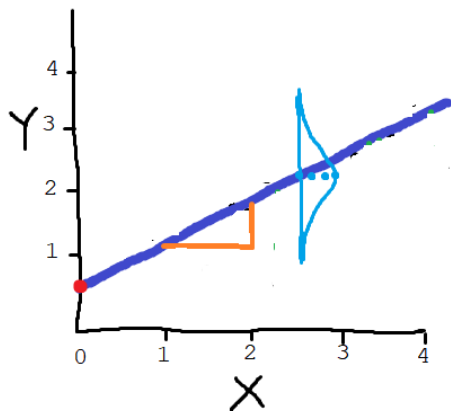
$$y = \alpha + \beta x + \epsilon$$

# Classification



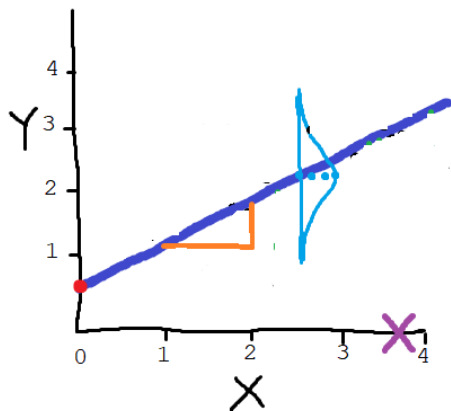
$$y = \alpha + \beta x + \epsilon$$
$$\epsilon \sim N(0, \sigma^2)$$

# Classification



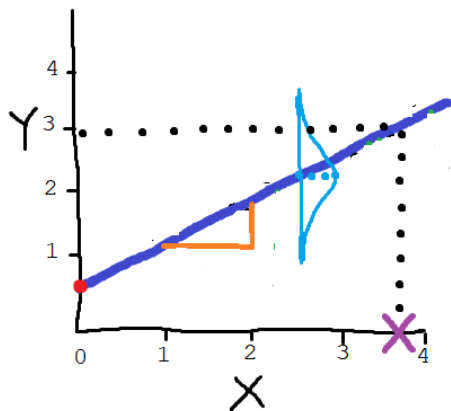
$$y = \alpha + \beta x + \epsilon$$
$$\epsilon \sim N(0, \sigma^2)$$

# Classification



$$y = \alpha + \beta x + \epsilon$$
$$\epsilon \sim N(0, \sigma^2)$$

# Classification

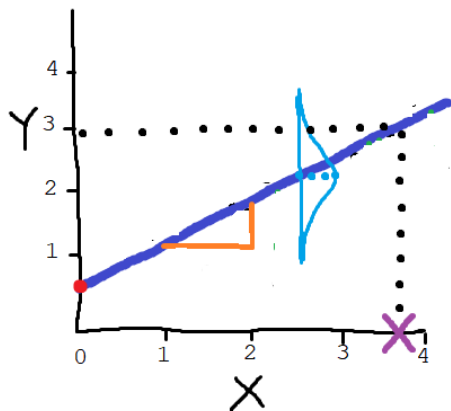


$$y = \alpha + \beta x + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

$$\hat{y} = \alpha + \beta x$$

# Classification



$$y = \hat{\alpha} + \hat{\beta}x + \epsilon$$
$$\epsilon \sim N(0, \hat{\sigma}^2)$$
$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

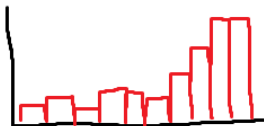
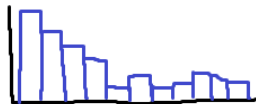
# Classification

- Training data (how much?, sampling?, pre-labeled?)
- Model (many, many options)
- Cross-validation (to avoid over-fitting)
- Best predictions are all that matter.



# Classification: Naive Bayes Classifier

# Classification: Naive Bayes Classifier



# Classification: Naive Bayes Classifier

Goal: estimate the probability that unknown document  $d$  belongs to the class ( $A$ ).

$$P(A|d) = \frac{P(d|A) P(A)}{P(d)}$$

The probability of the words given the class are denoted  $P(w_i|A)$ . Thus we take  $P(d|A)$  to be the independent product over all words in the document.

$$p(d|A) = \prod_i P(w_i|A)$$
$$P(A|d) = \frac{P(A)}{P(d)} \prod_i P(w_i|A)$$

We then assume that a document should be classified as either  $A$  or  $B$  and thus we can generate a symmetrical equation for  $B$  which can be used to produce a likelihood ratio:

$$P(B|d) = \frac{P(B)}{P(d)} \prod_i P(w_i|B)$$
$$\frac{P(A|d)}{P(B|d)} = \frac{P(A)}{P(B)} \prod_i \frac{p(w_i|A)}{p(w_i|B)}$$

Because in practice the product is quite small so we work with the log ratio.

$$\log \frac{P(A|d)}{P(B|d)} = \log \frac{P(A)}{P(B)} + \sum_i \log \frac{p(w_i|A)}{p(w_i|B)}$$

We estimate  $p(w_i|A)$  as the percentage of word  $w_i$  in document  $A$ . We denote this final quantity the score. Scores above 0 are in class  $A$ .

# Classification

R Exercise. Code: <http://www.mit.edu/~rnielsen/arabicTextWorkshop.zip>

Slides: [http://www.mit.edu/~rnielsen/arabic\\_text\\_slides.pdf](http://www.mit.edu/~rnielsen/arabic_text_slides.pdf)  
<http://www.mit.edu/~rnielsen/> > “Helpful Stuff” (at the bottom)

# Workshop Overview

1. Principles of text analysis research.
2. Acquiring text.
3. Pre-processesing text for quantitative analysis.
4. Topic Modeling (Unsupervised machine learning).
5. Classification (Supervised machine learning).
6. **Visualization.**

# Visualization



# Visualization

Goal: Communicate visually

# Visualization

## Principles:

- Every graph should have a point
- Clear labels
- Every element should serve a purpose
- Avoid redundant information
- Avoid misleading

# Visualization

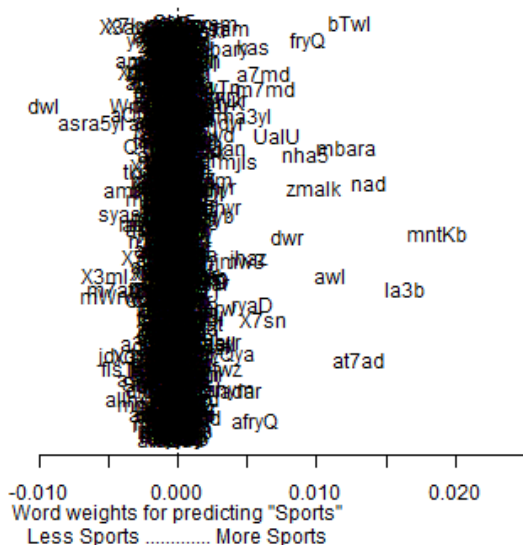
R Exercise. Code: <http://www.mit.edu/~rnielsen/arabicTextWorkshop.zip>

Slides: [http://www.mit.edu/~rnielsen/arabic\\_text\\_slides.pdf](http://www.mit.edu/~rnielsen/arabic_text_slides.pdf)  
<http://www.mit.edu/~rnielsen/> > “Helpful Stuff” (at the bottom)



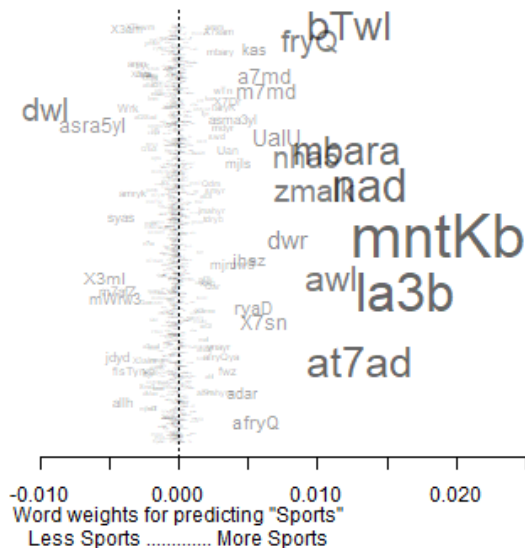
# Visualization: Naive Bayes Classifier

Which words predict Sports articles?



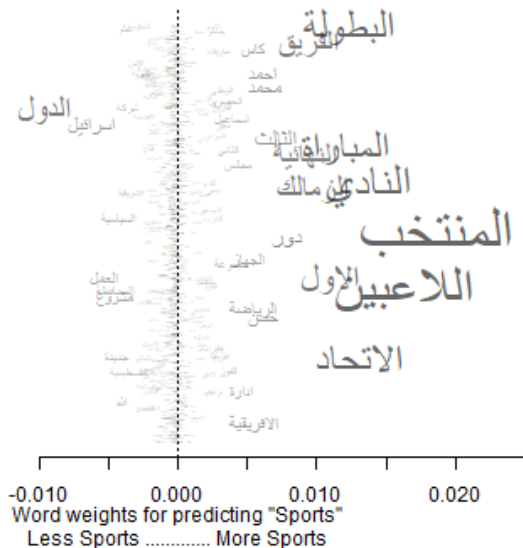
## Visualization: Naive Bayes Classifier

### Which words predict Sports articles?



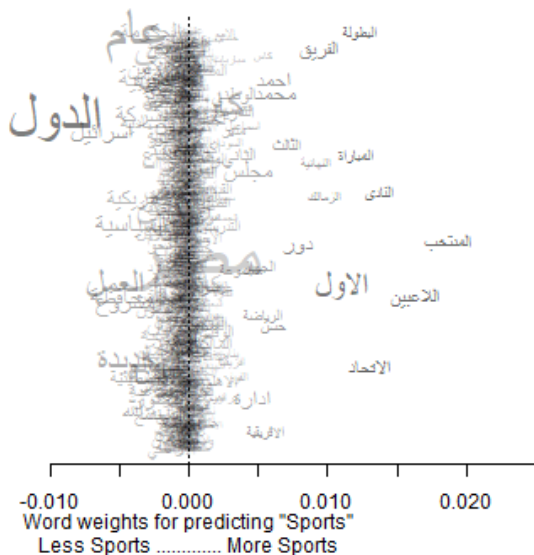
## Visualization: Naive Bayes Classifier

### Which words predict Sports articles?



# Visualization: Naive Bayes Classifier

Which words predict Sports articles?





## Visualization: Naive Bayes Classifier

