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/` > "Helpful Stuff" (at the bottom)

Introduction

In this session:

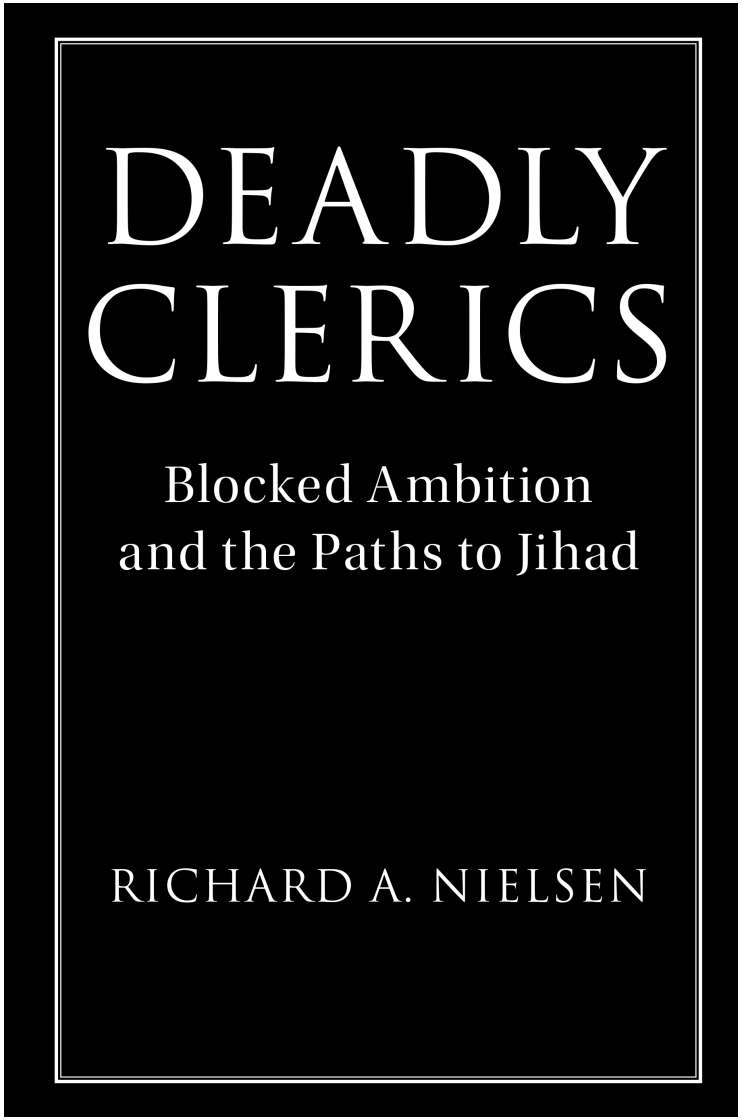
- Examples of text analysis
- Principles of text analysis research

Introduction

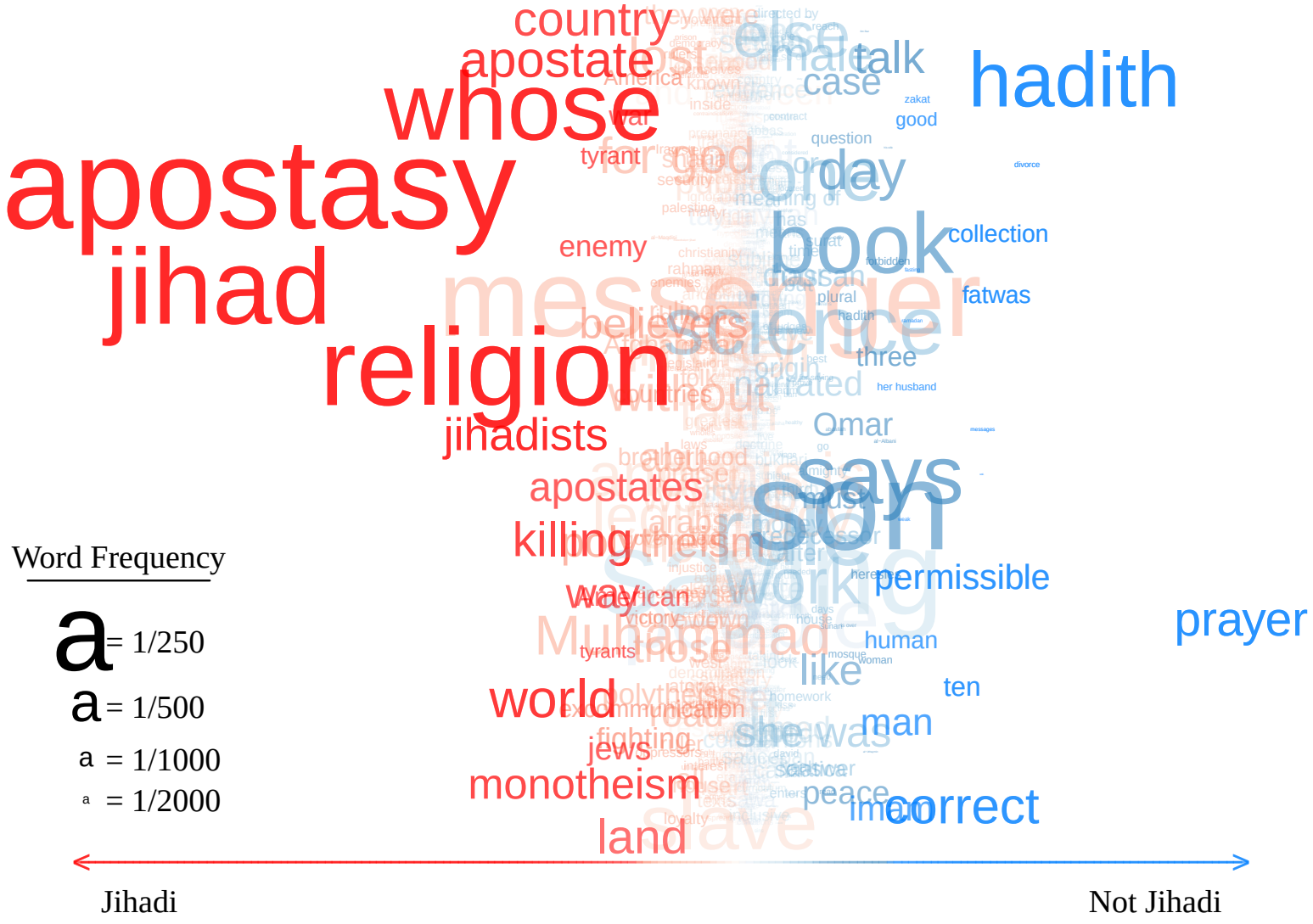
What is quantitative text analysis?

How do I use it in my research?

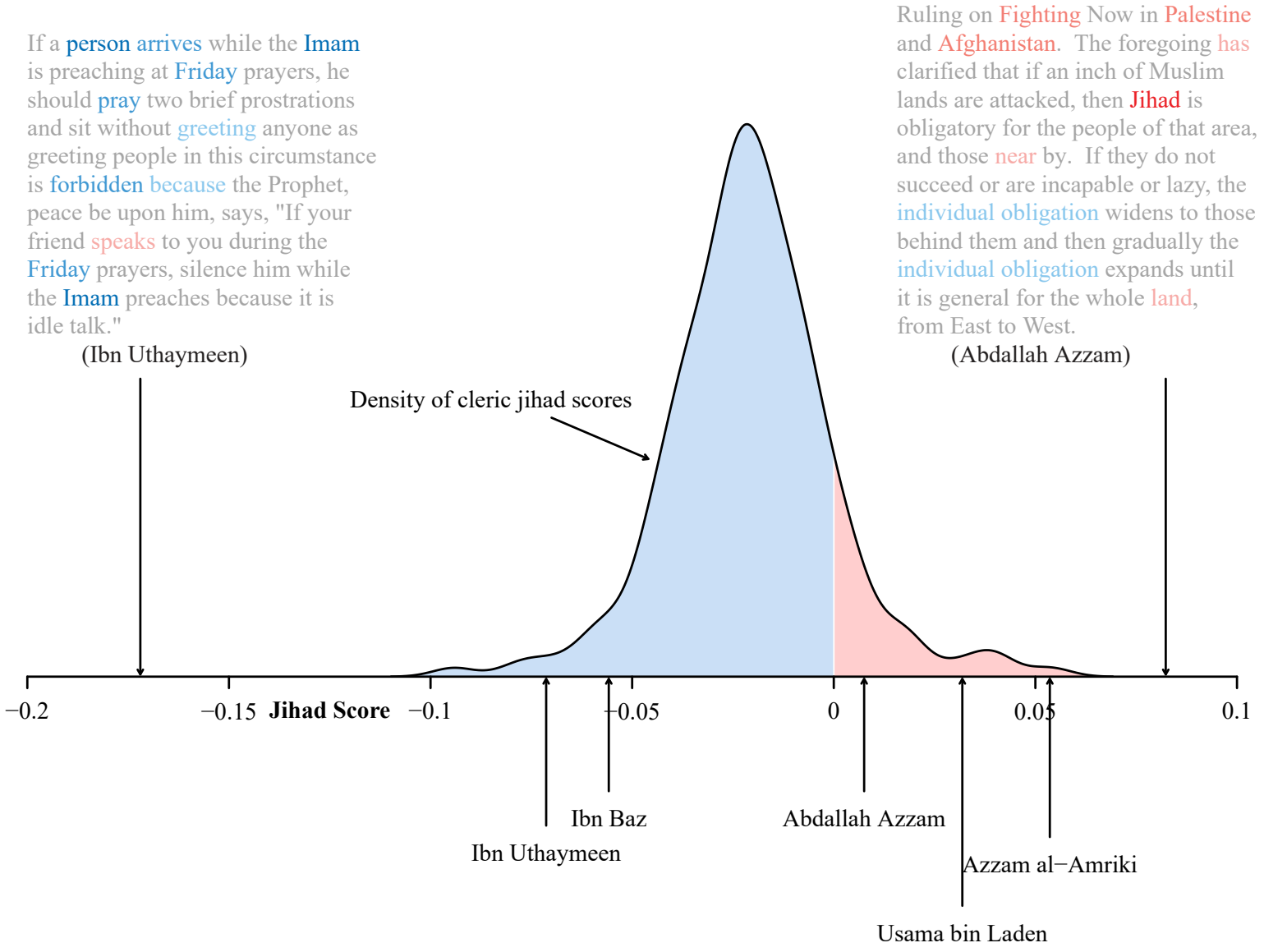
Introduction



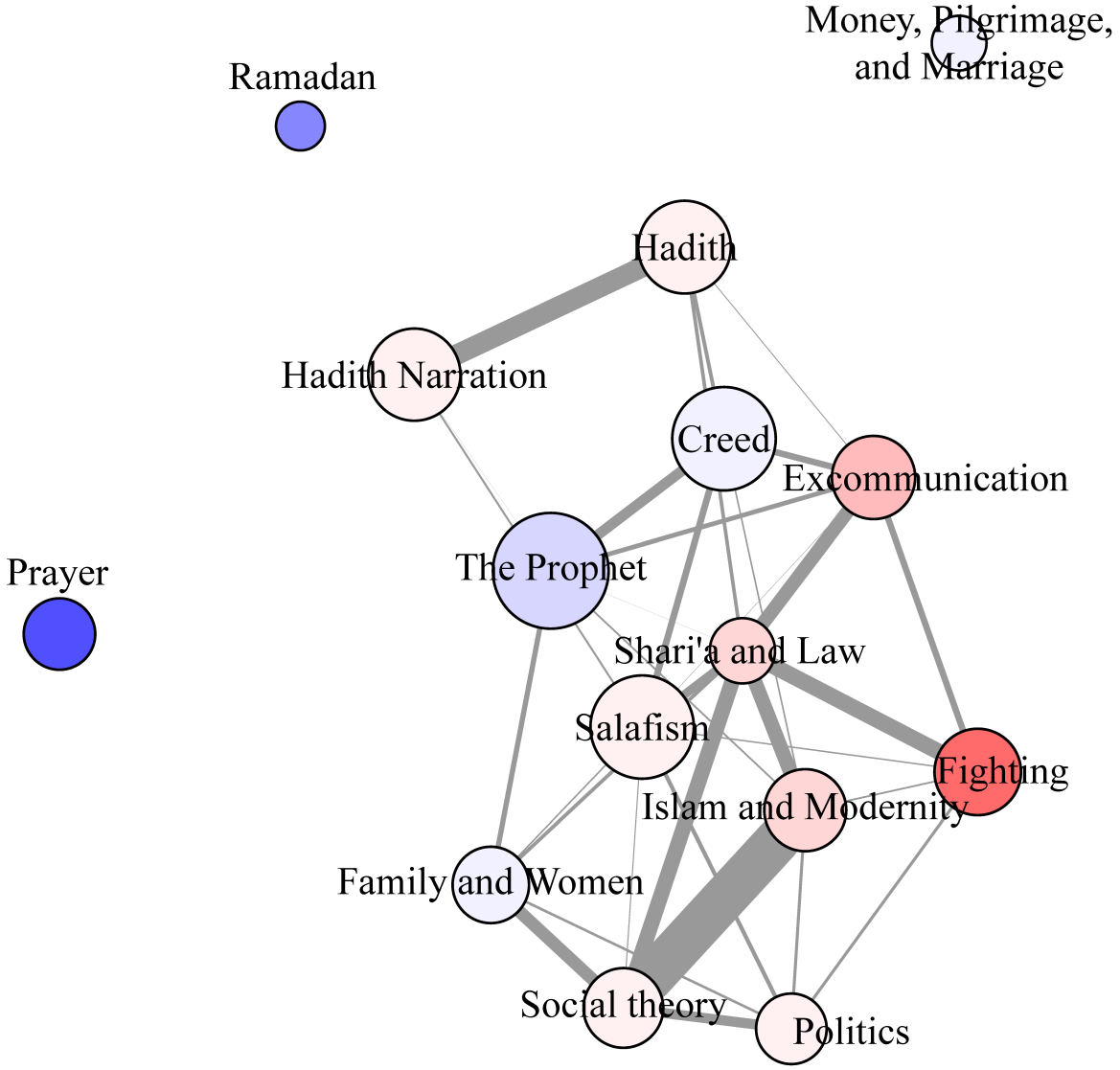
Introduction



Introduction



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. 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.

صيد الفوائد الرئيسة
- اعرف 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 - تغريدات
- واحة الأدب
- منوعات - مختارات
- الملل والنحل
- الطبيب الداعية
- بحوث علمية
- تربية الأبناء
- سيادة الشريعة

جديد مكتبة صيد الفوائد
المعلم الداعية

تغريدات بواسطة @saaidnet1

موقع صيد الفوائد @saaidnet1
بين يدي الأمير!
أحمد بن عبد المحسن العساف
@ahmalassaf

إدراج | مشاهدة على تويتر

صيد الفوائد مختارات

- الفروق المنهجيَّة بين أهل السُنَّة وأهل البدعة
- بين يدي الأمير!
- عشرةُ أبياتٍ في معنى
- القرآن : إعجاز علمي
- التدوين: الكلمة حين تنتصر!

صيد الفوائد الأنشطة الدعوية

صيد الفوائد جديد المكتبة
- 53 قاعدة لحياة أفضل
- قراءة نقدية لكتاب (ما بعد الصحوة) للأستاذ عبدالله الغذامي ...!
- قلوب تهوى العطاء
- 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

صيد الفوائد الرئيسة
- اعرف 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 - تغريدات
- واحة الأدب
- منوعات - مختارات
- الملل والنحل
- الطبيب الداعية
- بحوث علمية
- تربية الأبناء
- سيادة الشريعة

جديد مكتبة صيد الفوائد

تغريدات

بواسطة @saaidnet1

موقع صيد الفوائد @saaidnet1

بين يدي الأمير!

أحمد بن عبد المحسن العساف

@ahmalassaf

إدراج | مشاهدة على تويتر

صيد الفوائد مختارات

الفروق المنهجيَّة بين أهل السُنَّة وأهل البدعة

بين يدي الأمير!

عشرةُ أبياتٍ في معنى

القرآن : إعجاز علمي

التدوين: الكلمة حين تنتصر!

صيد الفوائد الأنشطة الدعوية

صيد الفوائد جديد المكتبة
- 53 قاعدة لحياة أفضل
- قراءة نقدية لكتاب (ما بعد الصحوة) للأستاذ عبدالله الغذامي ...!
- قلوب تهوى العطاء
- 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

ملتقى العلماء وطلبة العلم

موقع الشيخ عبدالعزيز بن بازرحمه الله كتبه	موقع الشيخ محمد العثيمين رحمه الله كتبه
موقع الشيخ عبدالله الجبرين كتبه	موقع الشيخ صالح بن فوزان الفوزان كتبه
الشيخ محمد ناصر الدين الألباني	موقع الشيخ عبدالرحمن السعدي كتبه
الشيخ عبدالرحمن بن ناصر البراك	
الشيخ فيصل بن عبدالعزيز آل مبارك	الشيخ عَبدِ الْعَزِيزِ بْنِ مُحَمَّدٍ السَّلمانِ
العلامة عبد الرحمن بن محمد بن	عبد الله بن جارالله آل جارالله رحمه

صيد الفوائد الرئيسة
- اعرف 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 - تغريدات
- واحة الأدب
- منوعات - مختارات
- الملل والنحل
- الطبيب الداعية
- بحوث علمية
- تربية الأبناء
- سيادة الشريعة

صيد الفوائد ملتقى طلبة العلم
رابط 1
رابط 2
رابط 3
رابط 4

صيد الفوائد مواقع اسلامية

ملتقى العلماء وطلبة العلم

موقع الشيخ عبدالعزيز بن بازرحمه الله كتبه	موقع الشيخ محمد العثيمين رحمه الله كتبه
موقع الشيخ عبدالله الجبرين كتبه	موقع الشيخ صالح بن فوزان الفوزان كتبه
الشيخ محمد ناصر الدين الألباني	موقع الشيخ عبدالرحمن السعدي كتبه
الشيخ عبدالرحمن بن ناصر البراك	
الشيخ فيصل بن عبدالعزيز آل مبارك	الشيخ عَبدِ الْعَزِيزِ بْنِ مُحَمَّدٍ السَّلمانِ
العلامة عبد الرحمن بن محمد بن	عبد الله بن جارالله آل جارالله رحمه

صيد الفوائد الرئيسة
- اعرف 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 - تغريدات
- واحة الأدب
- منوعات - مختارات
- الملل والنحل
- الطبيب الداعية
- بحوث علمية
- تربية الأبناء
- سيادة الشريعة

صيد الفوائد ملتقى طلبة العلم
رابط 1
رابط 2
رابط 3
رابط 4

صيد الفوائد مواقع اسلامية

صيد الفوائد الرئيسة
- اعرف 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 - تغريدات
- واحة الأدب
- منوعات - مختارات
- الملل والنحل
- الطبيب الداعية
- بحوث علمية
- تربية الأبناء
- سيادة الشريعة

صيد الفوائد ملتقى الداعيات
- ملتقى الداعيات
- د.أميمة الجلاهمة
- أفراح الحميضي
- رقية المحارب
- د.بنت الرسالة
- أ.مها 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 لبّان
- صاحبة قلم
- لبنى شرف

صيد الفوائد الرئيسة
- اعرف 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 - تغريدات
- واحة الأدب
- منوعات - مختارات
- الملل والنحل
- الطبيب الداعية
- بحوث علمية
- تربية الأبناء
- سيادة الشريعة

صيد الفوائد ملتقى الداعيات
- ملتقى الداعيات
- د.أميمة الجلاهمة
- أفراح الحميضي
- رقية المحارب
- د.بنت الرسالة
- أ.مها 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 لبّان
- صاحبة قلم
- لبنى شرف

صيد الفوائد الرئيسة
- اعرف 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 - تغريدات
- واحة الأدب
- منوعات - مختارات
- الملل والنحل
- الطبيب الداعية
- بحوث علمية
- تربية الأبناء
- سيادة الشريعة

صفحة الدكتورة أميمة بنت أحمد الجلاهمة

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

صيد الفوائد أميمة الجلاهمة
- مقالات
- الصفحة الرئيسية
- ملتقى الداعيات
- للنساء فقط

صيد الفوائد مواقع اسلامية

صيد الفوائد الرئيسة
- اعرف 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 - تغريدات
- واحة الأدب
- منوعات - مختارات
- الملل والنحل
- الطبيب الداعية
- بحوث علمية
- تربية الأبناء
- سيادة الشريعة

صفحة الدكتورة أميمة بنت أحمد الجلاهمة

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

صيد الفوائد أميمة الجلاهمة
- مقالات
- الصفحة الرئيسية
- ملتقى الداعيات
- للنساء فقط

صيد الفوائد مواقع اسلامية

TWEETS 19.6K | FOLLOWING 256 | FOLLOWERS 19.7K | LIKES 5,454

أميمة أحمد الجلاهمة

@OmaimaAlJalahma

كاتبة و أستاذ مشارك في جامعة الدمام، تخصص عقيدة، مقارنة الأديان

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 دخلَ بيتي مُؤْمِنًا وَلِلْمُؤْمِنِينَ وَالْمُؤْمِنَاتِ وَلَا تَزِدِ الظَّالِمِينَ إِلَّا تَبَارًا)

45 25

أميمة أحمد الجلاهمة Retweeted

@asuwayed عبدالعزيز السويد · 23h

دي ميستورا: ما يجري في مدينة #حلب يرقى إلى جرائم حرب #روسيا #سورنا

10

أميمة أحمد الجلاهمة Retweeted

@Naifchair كرسي الأمير نايف · Sep 25

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

أميمة أحمد الجلاهمة

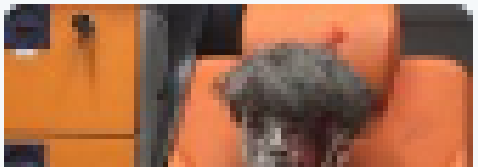
@OmaimaAlJalahma

كاتبة و أستاذ مشارك في جامعة الدمام، تخصص عقيدة، مقارنة الأديان

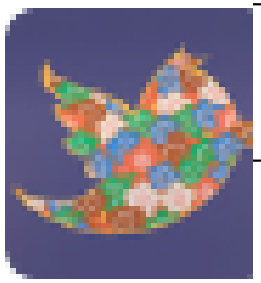
Saudi Arabia

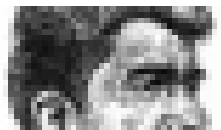
Joined October 2011

52 Photos and videos



Twee





صيد الفوائد الرئيسة
- اعرف 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 - تغريدات
- واحة الأدب
- منوعات - مختارات
- الملل والنحل
- الطبيب الداعية
- بحوث علمية
- تربية الأبناء
- سيادة الشريعة

صيد الفوائد ملتقى الداعيات
- ملتقى الداعيات
- د.أميمة الجلاهمة
- أفراح الحميضي
- رقية المحارب
- د.بنت الرسالة
- أ.مها 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 لبّان
- صاحبة قلم
- لبنى شرف

صيد الفوائد الرئيسة

- اعرف 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 - تغريدات
- واحة الأدب
- منوعات - مختارات
- الملل والنحل
- الطبيب الداعية
- بحوث علمية
- تربية الأبناء
- سيادة الشريعة

صيد الفوائد ملتقى الداعيات

- ملتقى الداعيات
- د.أميمة الجلاهمة
- أفراح الحميضي
- رقية المحارب
- د.بنت الرسالة
- أ.مها 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 لبّان
- صاحبة قلم
- لبنى شرف

صفحة الدكتورة افراح بنت علي الحميضي

أورام البناء الثقافي
بيدين، بكتفين، بخطوتين معاً
للنساء: جرعة فقه تقي من غوائل الجهل
هوس نسائي..!
لمـــاذا .. 8 مارس؟؟
معلِّمة بشهادة جلاد!!
محصّلة التشبُّه: عرى الأجساد، خواء العقول!!
الطلاق ولعبة النَّار
Power Rangers والبقية ستأتي..
أصواتٌ تطرب لها الذئاب!!
ما لم يدركه الرَّقيب

الآن
الموقع باللغة الإنجليزية
الأسهم والمعاملات المالية

صفحة الدكتورة افراح بنت علي الحميضي

أورام البناء الثقافي
بيدين، بكتفين، بخطوتين معاً
للنساء: جرعة فقه تقي من غوائل الجهل
هوس نسائي..!
لمـــاذا .. 8 مارس؟؟
معلِّمة بشهادة جلاد!!
محصّلة التشبُّه: عرى الأجساد، خواء العقول!!
الطلاق ولعبة النَّار
Power Rangers والبقية ستأتي..
أصواتٌ تطرب لها الذئاب!!
ما لم يدركه الرَّقيب

الآن
الموقع باللغة الإنجليزية
الأسهم والمعاملات المالية

Power Rangers والبقية ستأتي..

د.أفراح بنت علي الحميضي

(Power Rangers) فيلم أجنبي مُترجم يخدم فئة الطفولة المبكّرة 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 وتتلقّفه أيدي النشء ليصيغ في نفوسهم - هو وأشباهه الكثير - نماذج سيئة وقيما مدمّرة؛ أقلّها تمجيد الإنسان الغربي ومنحه قوة تنافس، بل تتفوّق على قوة الله العظيم - أستغفر الله -.

تتمحور قصّة 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.

صيد الفوائد مواقع اسلامية

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...

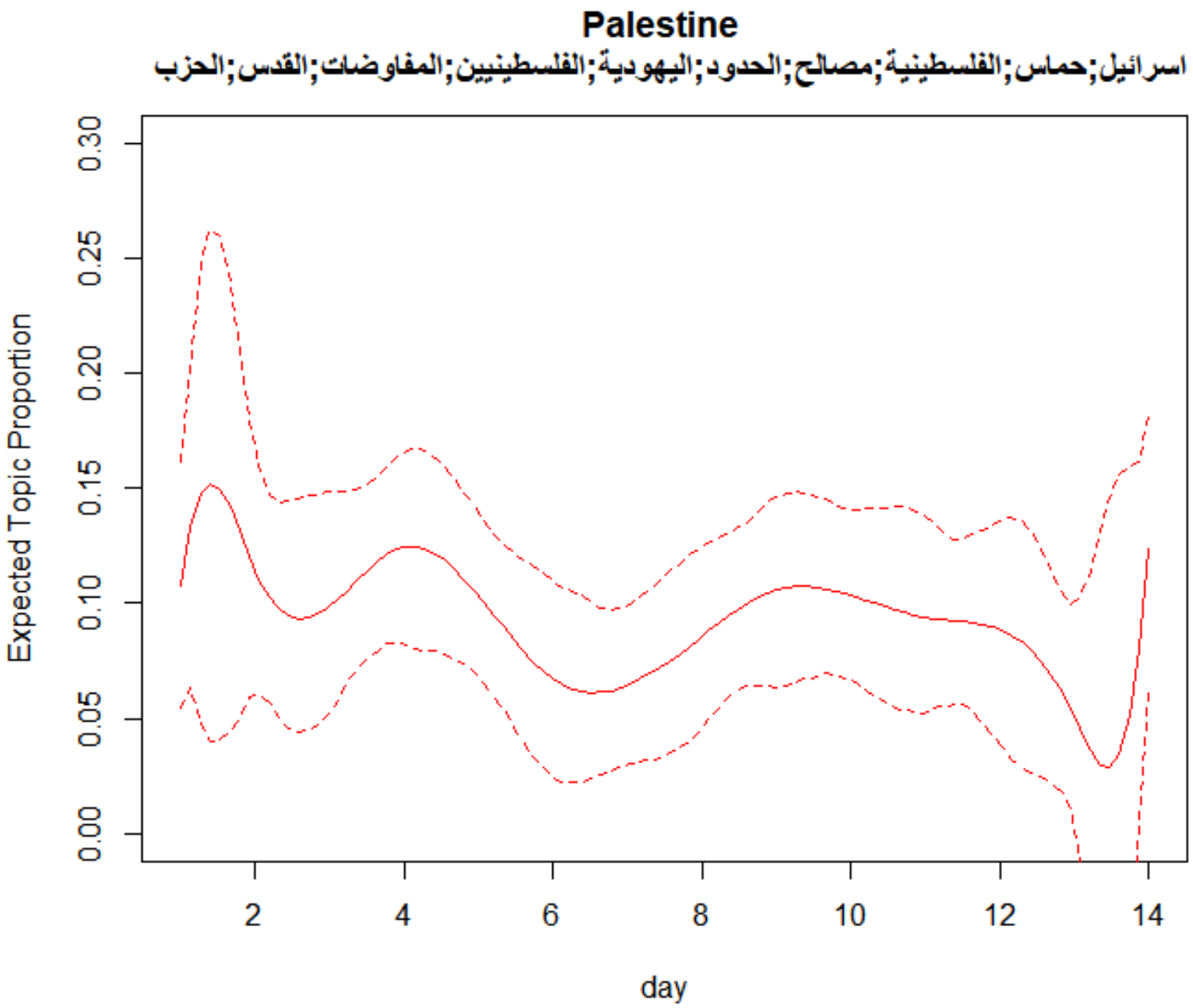
Workshop Overview

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

Workshop Overview

Workshop Overview

Workshop Overview



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)

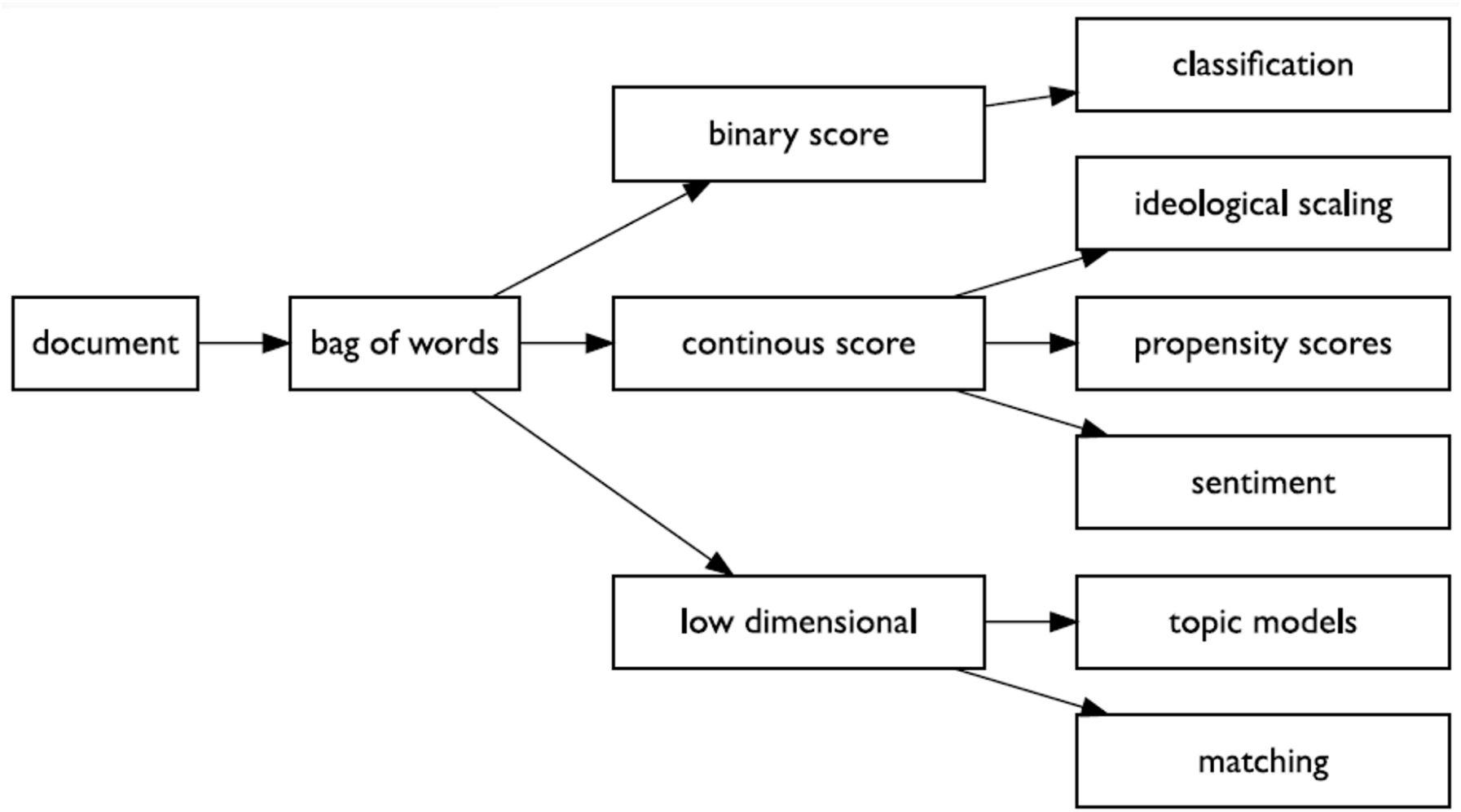
(1) All quantitative models of language are wrong—but some are useful.
(2) Quantitative methods for text amplify resources and augment humans.
(3) There is no globally best method for automated text analysis.
(4) Validate, Validate, Validate.

My additions:

(5) All text analysis is dimension reduction.

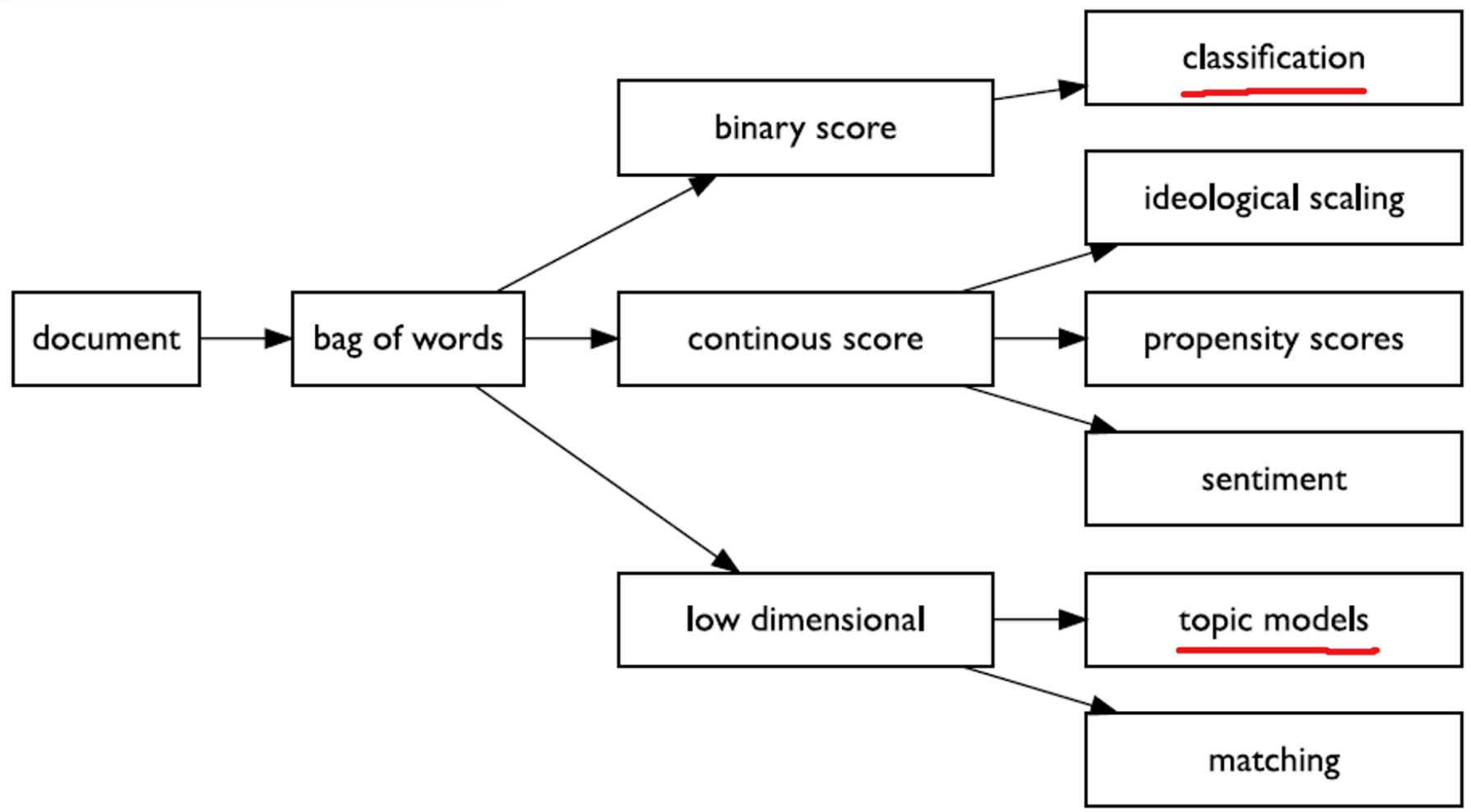
(6) There is no free lunch.
(supervised methods require front-end work, unsupervised require back-end work)

Principles of Text Analysis Research



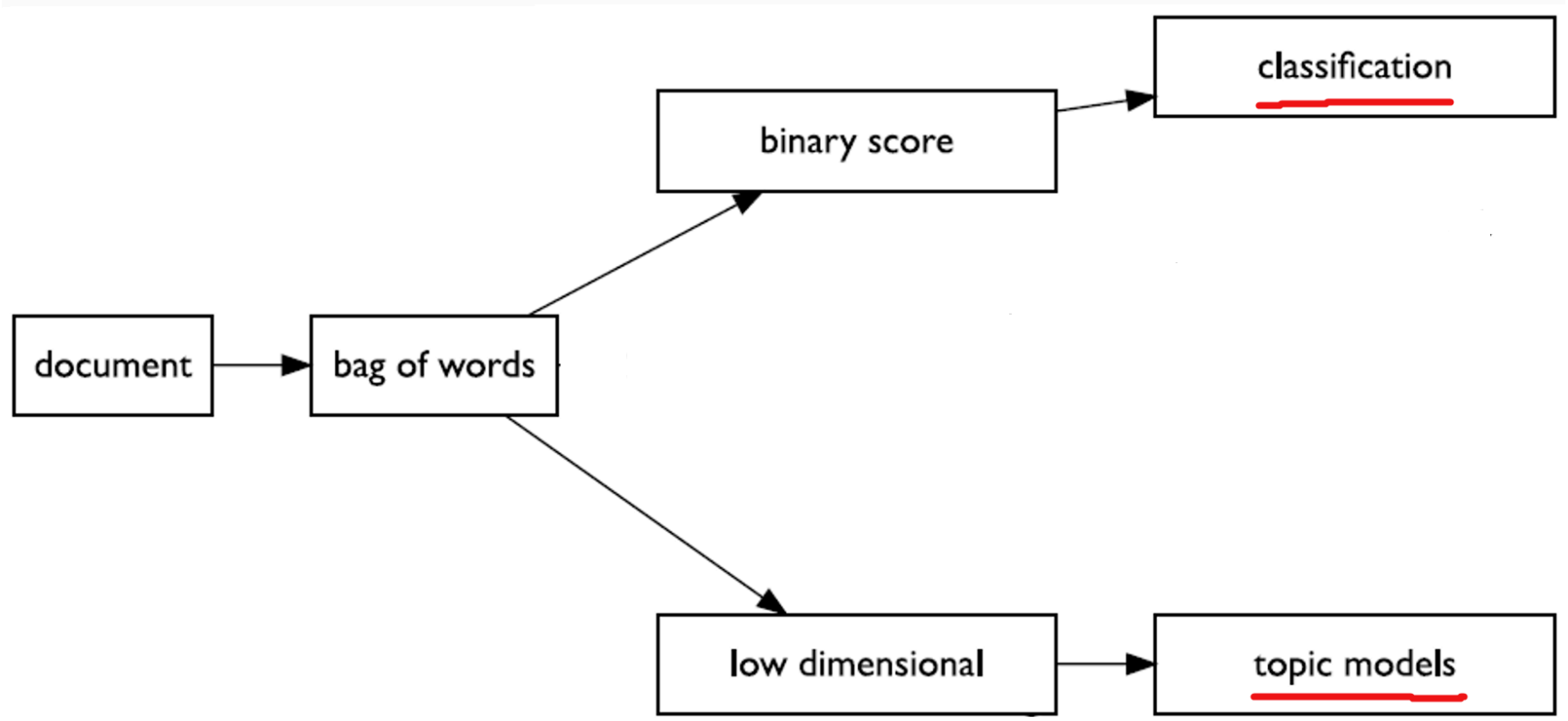
2nd generation text analysis

Principles of Text Analysis Research



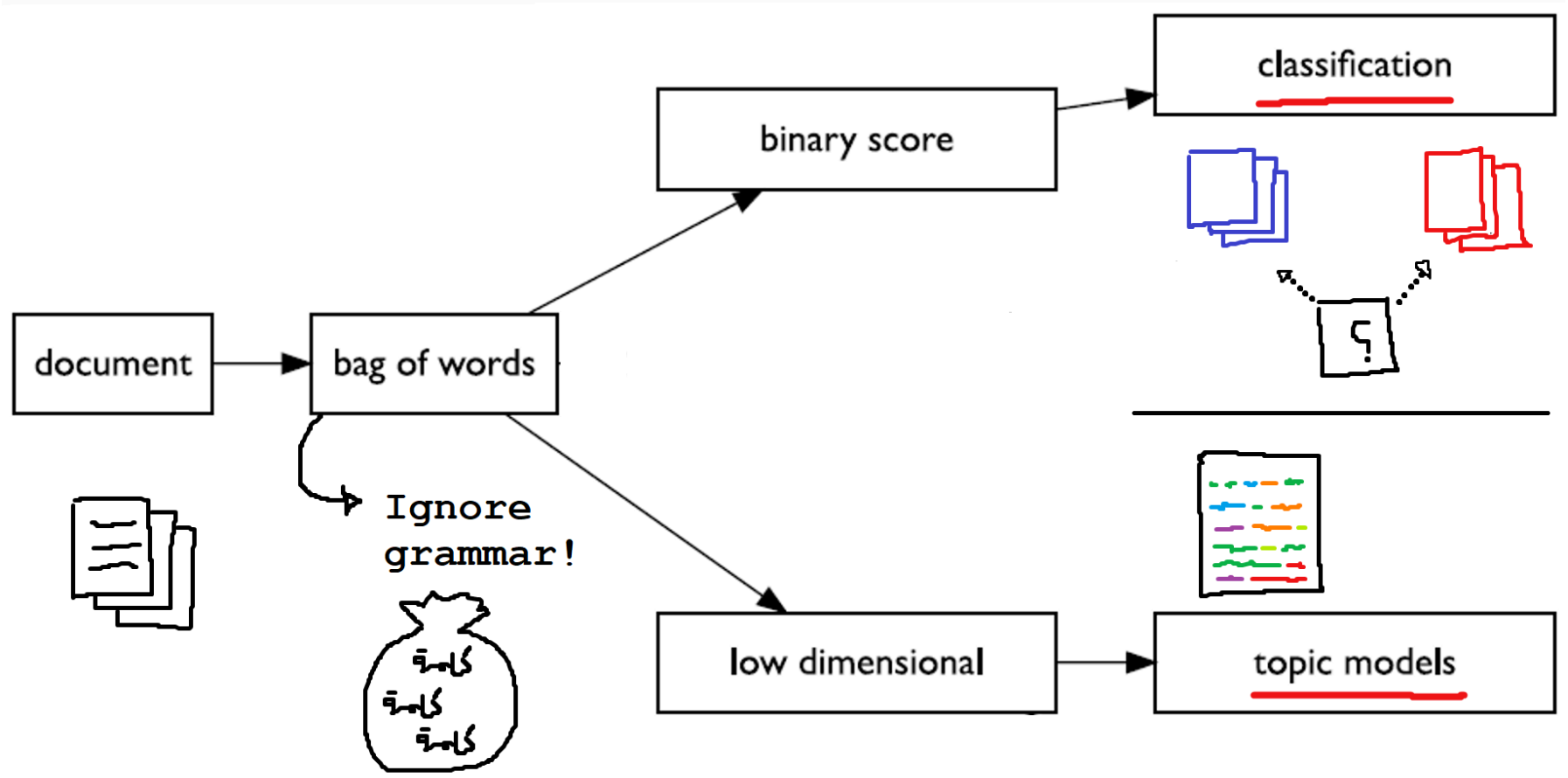
2nd generation text analysis

Principles of Text Analysis Research



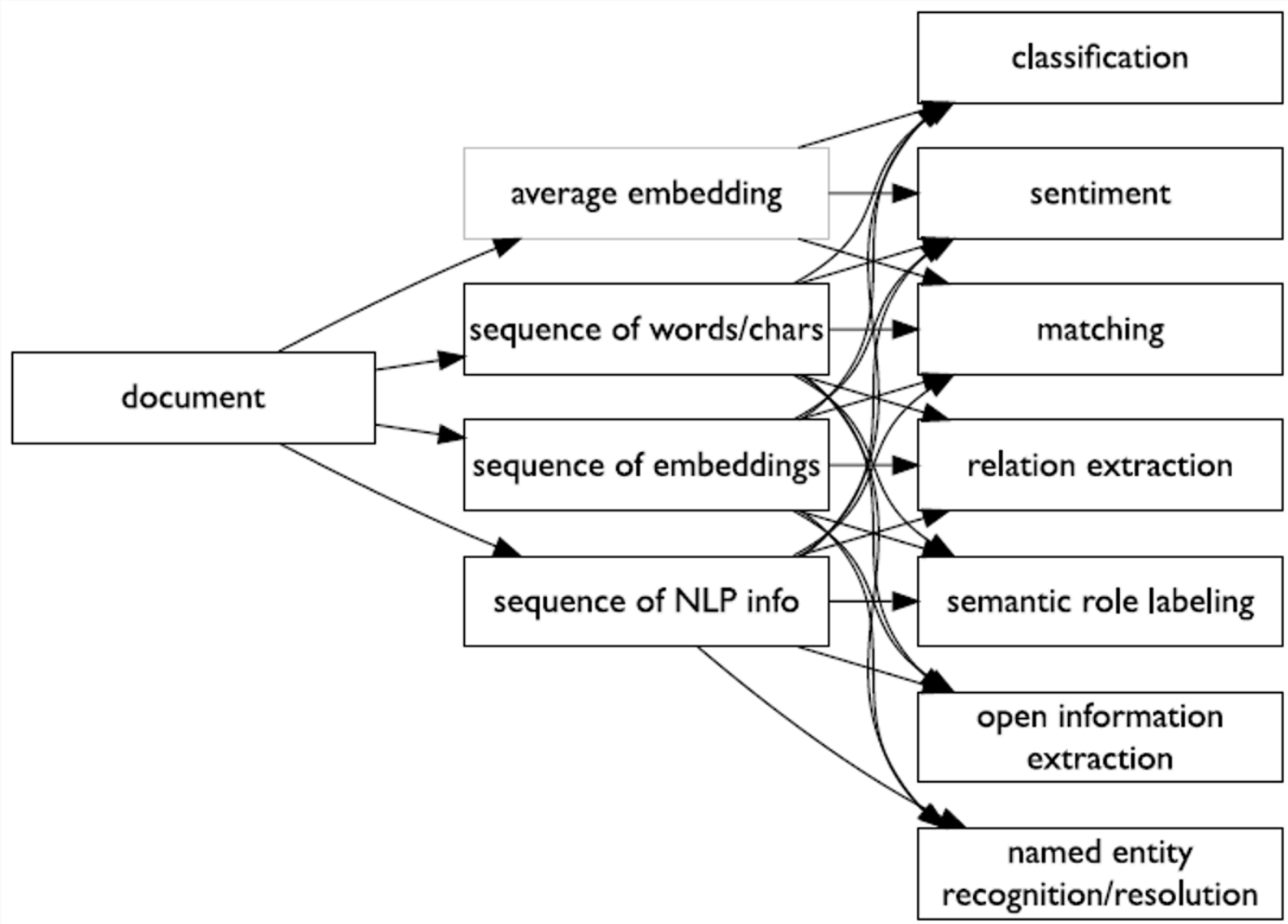
2nd generation text analysis

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

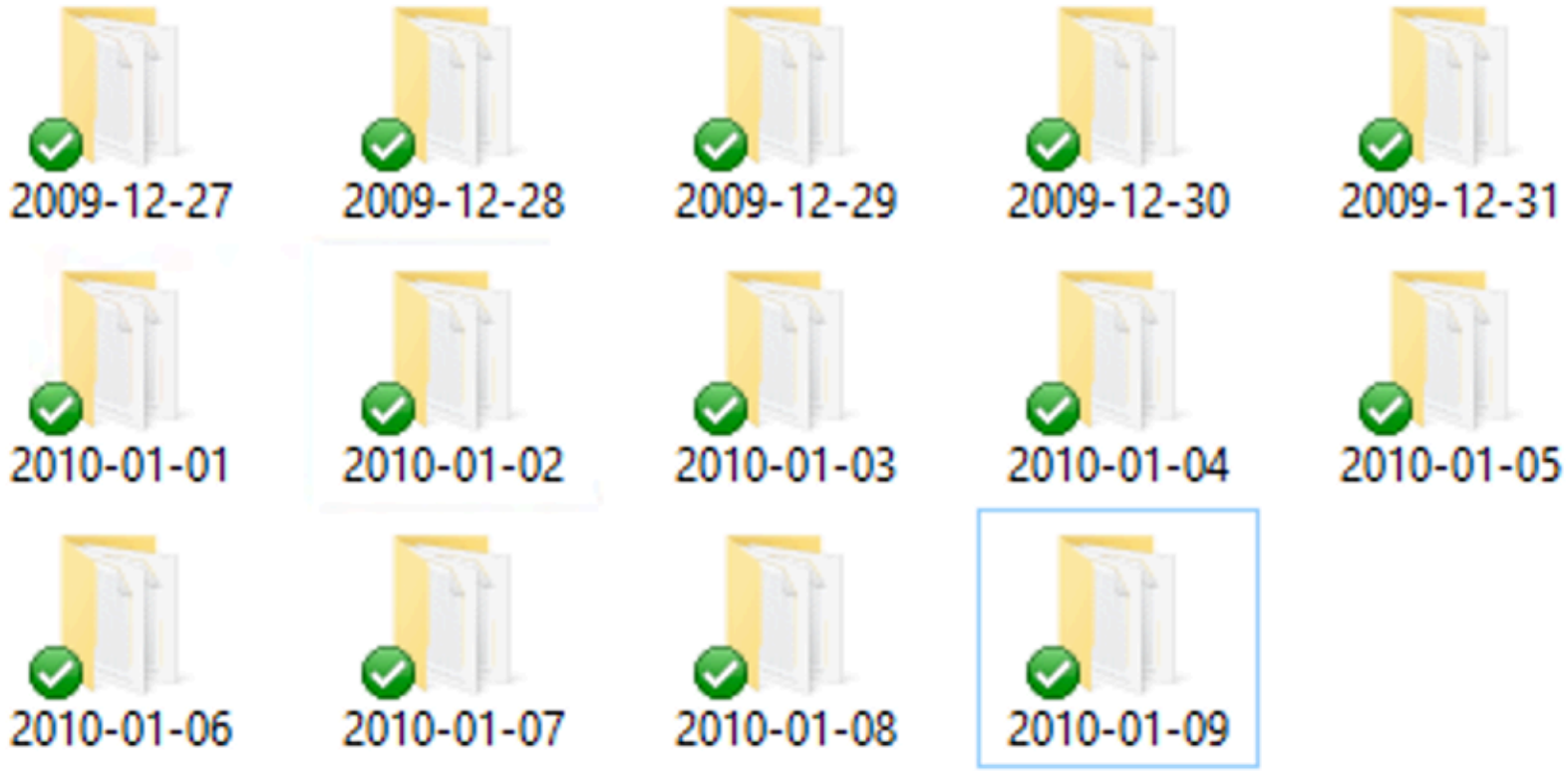
The goal: .txt files in a directory

Acquiring Text

The goal: .txt files in a directory

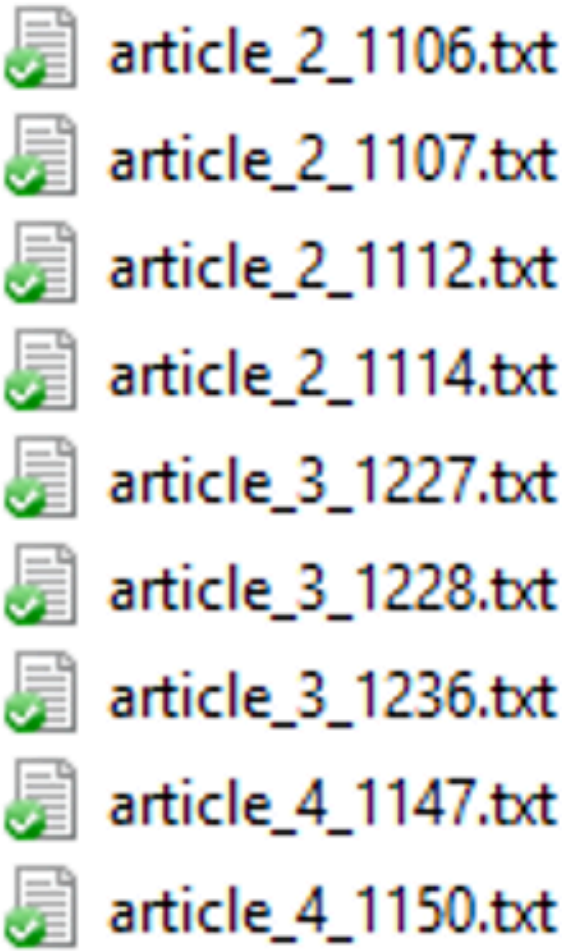
Acquiring Text

The goal: .txt files in a directory



Acquiring Text

The goal: .txt files in a directory



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 th
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/	Automatic redirection to servers worldwide, currently sponsorec
http://cloud.r-project.org/	Automatic redirection to servers worldwide, currently sponsorec
Algeria	
https://cran.usthb.dz/	University of Science and Technology Houari Boumediene
http://cran.usthb.dz/	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/	CSIRO
http://cran.csiro.au/	CSIRO
https://mirror.aarnet.edu.au/pub/CRAN/	AARNET
https://cran.ms.unimelb.edu.au/	School of Mathematics and Statistics, University of Melbourne
https://cran.curtin.edu.au/	Curtin University of Technology
Austria	
https://cran.wu.ac.at/	Wirtschaftsuniversität Wien
http://cran.wu.ac.at/	Wirtschaftsuniversität Wien
Belgium	
http://www.freestatistics.org/cran/	K.U.Leuven Association
https://lib.ugent.be/CRAN/	Ghent University Library
http://lib.ugent.be/CRAN/	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 lik
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
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 so
code. The sources have to be compiled before you can use them. If you do not know what this means, you probab
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 versio
- 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 feat
  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	Binaries for base distribution. This is what you want to **install R for the first time**.
contrib	Binaries of contributed CRAN packages (for R >= 2.13.x; managed by Uwe Ligges). There is also informati on third party software available for CRAN Windows services and corresponding environment and make variables.
old contrib	Binaries of contributed CRAN packages for outdated versions of R (for R < 2.13.x; managed by Uwe Ligges
Rtools	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 / suggestio 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.

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 Ma 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 s (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 bi 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 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 su 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 proce 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 signa

```
pkgutil --check-signature R-3.4.2.pkg
```

Files:

R-3.4.2.pkg
MD5-hash: 71d2db804 38cf2ce8ff9cfed1fa860
SHA-hash: c577af8e 06e5a230b2932545340e9e1273dba
(ca. 61MB)

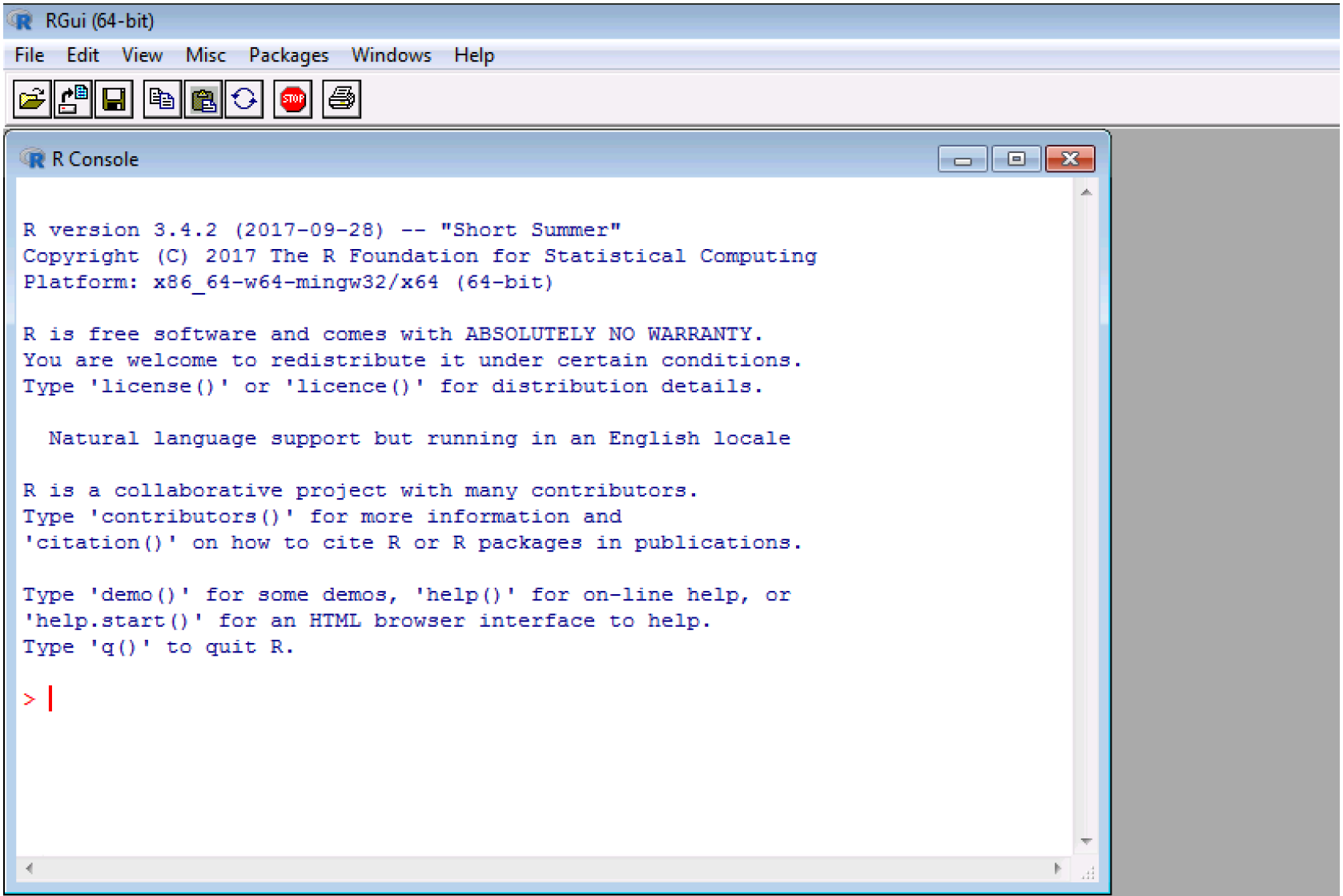
R 3.4.2 binary for OS X 10.11 (El Capitan) and higher, signed package. Contains R 3.4.2 framework, R.app GUI 1.70 in 64-bit for Intel Macs, Tcl/Tk 8.6.6 X11 libraries and Texinfo 5.2. The latter two components are optional and can be ommitted when choosing "custom install", they are only needed if you 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

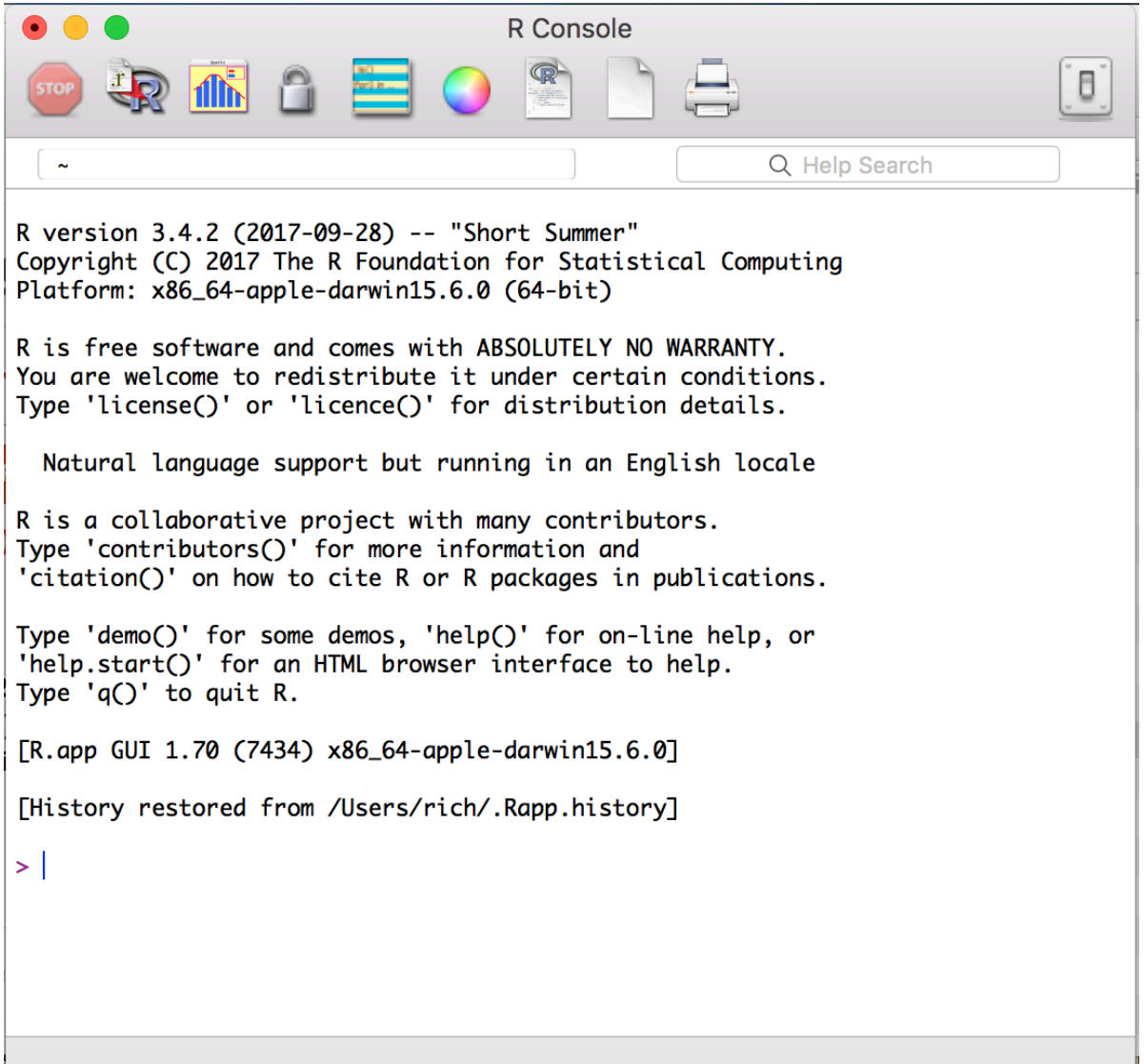
Acquiring Text: Installing R

R console: Windows



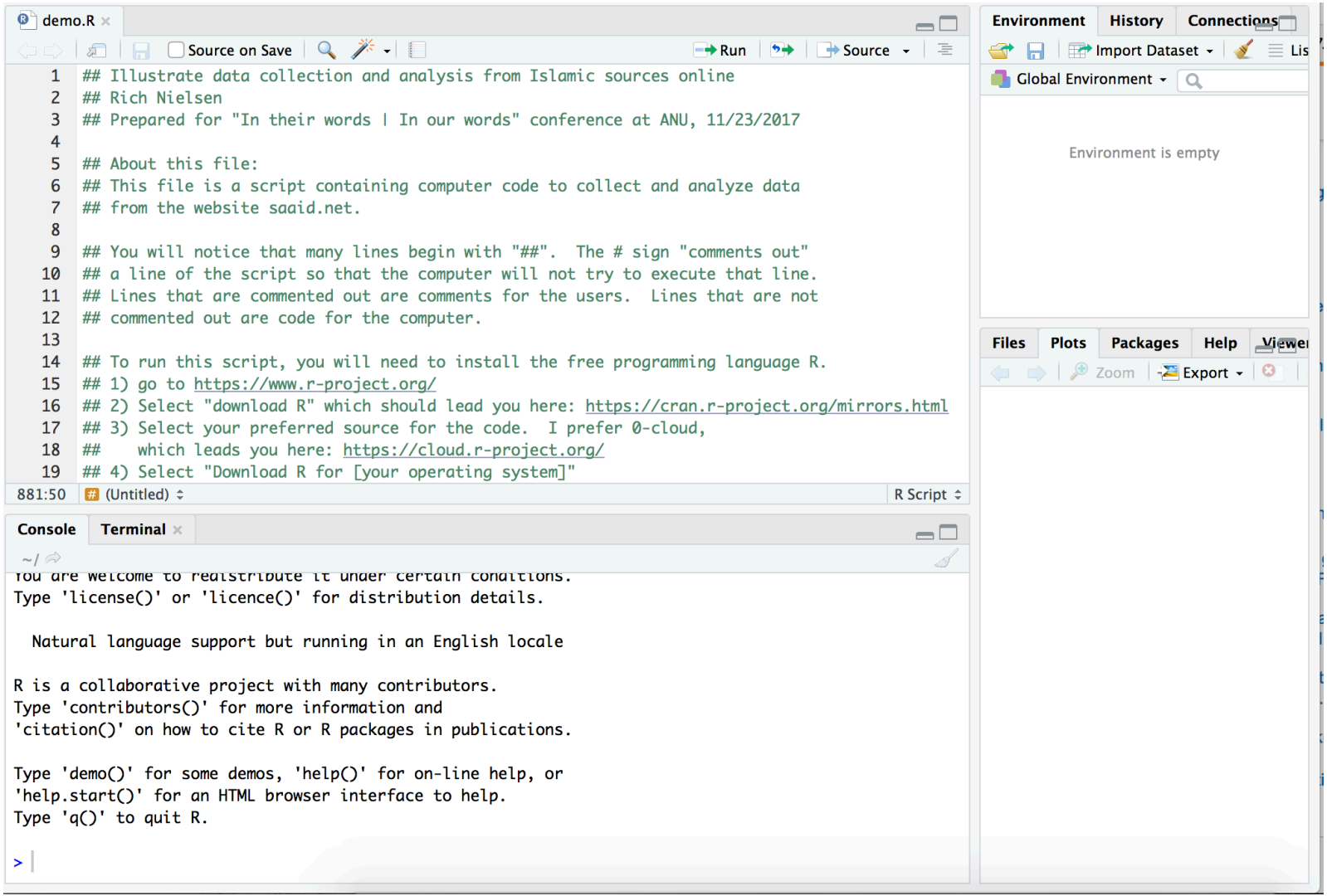
Acquiring Text: Installing R

R console: Mac



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/` > "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

```
    <div style="float: right">
        <a href="http://twitter.com/share" class="twitter-share-button" data-c
    </div>
    <div id="ReaderCount" class="innerb_main" style="float: right;"><div class
    <g:plusone></g:plusone>
</div>
<div id="abstractDiv">
    <img src="images/gry-sep-daily.jpg" alt="" style="margin-top: 10px;" />
    <div id="iqraa">
    </div>
    <br />
    <span id="txtBrief">مؤسس ساقية عبد المنعم الصاوي, الذي تحدث عن تجربة الساقية في الشارع المصري
</div>
<div id="NewsImage"><div id='divImages'><div class='floatimgleft'><img border=
<div id="MoreImages">
</div>
<div id="bodyDiv">
    <div id="txtBody" class="bbBodyp" align="justify"><p align='justify'><p>كر

</div>
```

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 .

Tokenize

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

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

صيد الفوائد الرئيسة
- اعرف 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 - تغريدات
- واحة الأدب
- منوعات - مختارات
- الملل والنحل
- الطبيب الداعية
- بحوث علمية
- تربية الأبناء
- سيادة الشريعة

صيد الفوائد ملتقى الداعيات
- ملتقى الداعيات
- د.أميمة الجلاهمة
- أفراح الحميضي
- رقية المحارب
- د.بنت الرسالة
- أ.مها 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 لبّان
- صاحبة قلم
- لبنى شرف

Term	% of documents using term		% of all words	
	Men	Women	Men	Women
allāh	95%	85%	3.0%	1.6%
ṣalā allāh	67	32	0.50	0.21
raḍī allāh	50	20	0.18	0.10
ḥadīth	64	39	0.30	0.14
qāl/yaqūl (al-)rasūl	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/` > "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.

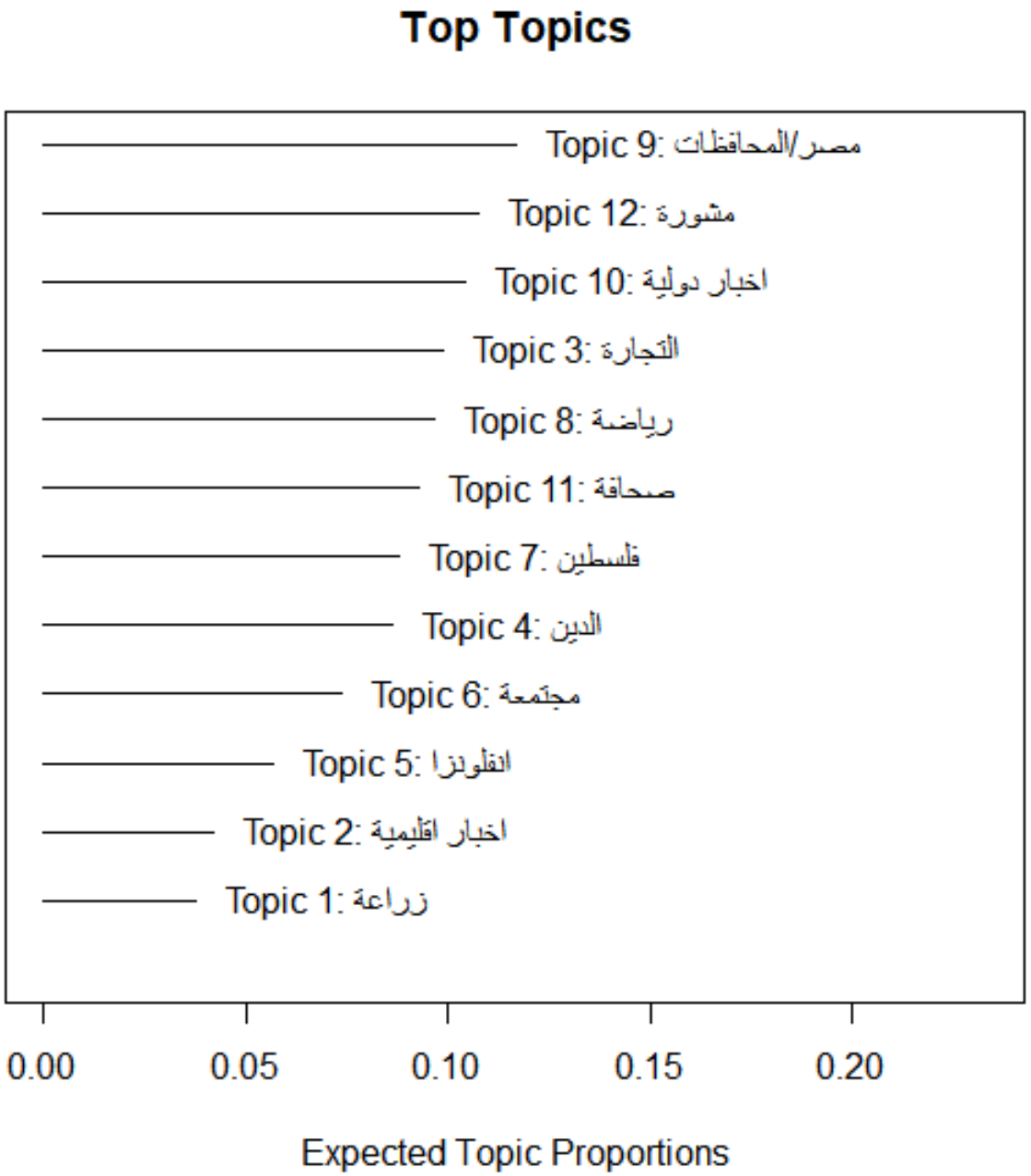
Good: low effort, quick

Bad: Not always sure what you get

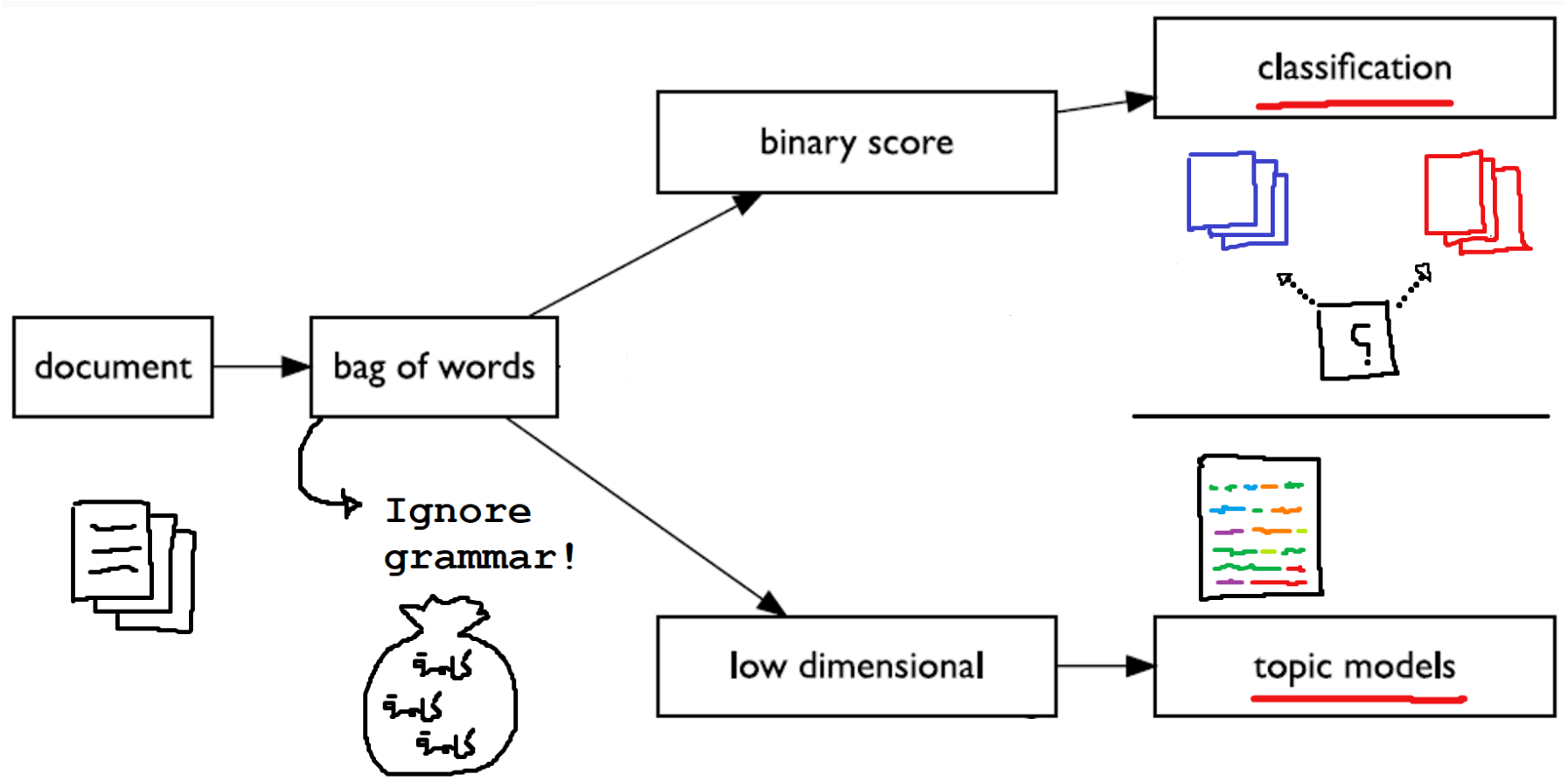
Topic Models

```
> 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
```

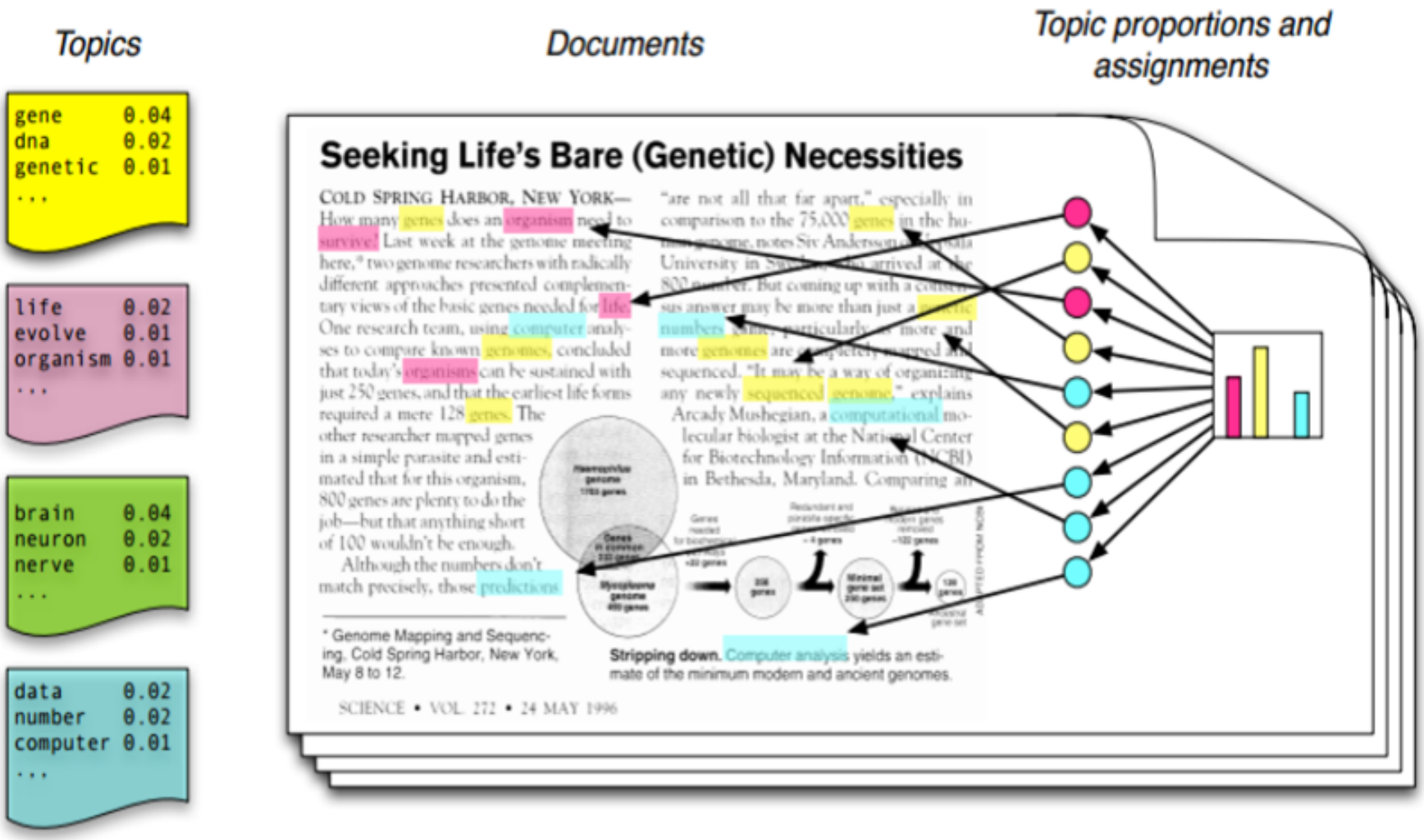
Topic Models



Topic Models



Topic Models



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 (DxK)

2) Topical Content Matrix (VxK)

Topic models

Two matrices estimated:

1) Topical Prevalence Matrix (DxK)

$$\boldsymbol{\theta} = \left[\begin{array}{c|cccc} & Topic1 & Topic2 & \dots & TopicK \\ \hline Doc1 & .2 & .1 & \dots & 0.05 \\ Doc2 & .2 & .1 & \dots & .3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ DocD & 0 & 0 & \dots & .5 \end{array}\right]$$

2) Topical Content Matrix (VxK)

Topic models

Two matrices estimated:

1) Topical Prevalence Matrix (DxK)

$$\boldsymbol{\theta} = \left[\begin{array}{c|cccc} & Topic1 & Topic2 & \dots & TopicK \\ \hline Doc1 & .2 & .1 & \dots & 0.05 \\ Doc2 & .2 & .1 & \dots & .3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ DocD & 0 & 0 & \dots & .5 \end{array}\right]$$

2) Topical Content Matrix (VxK)

$$\boldsymbol{\beta}^{\boldsymbol{T}} = \left[\begin{array}{c|cccc} & Topic1 & Topic2 & \dots & TopicK \\ \hline \text{“}text\text{”} & .02 & .001 & \dots & 0.001 \\ \text{“}data\text{”} & .001 & .02 & \dots & 0.001 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{“}analysis\text{”} & .01 & .01 & \dots & 0.0005 \end{array}\right]$$

Topic models

Two matrices estimated:

$X \approx \theta\beta$

1) Topical Prevalence Matrix (DxK)

$$\boldsymbol{\theta} = \left[\begin{array}{c|cccc} & Topic1 & Topic2 & \dots & TopicK \\ \hline Doc1 & .2 & .1 & \dots & 0.05 \\ Doc2 & .2 & .1 & \dots & .3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ DocD & 0 & 0 & \dots & .5 \end{array}\right]$$

2) Topical Content Matrix (VxK)

$$\boldsymbol{\beta^T} = \left[\begin{array}{c|cccc} & Topic1 & Topic2 & \dots & TopicK \\ \hline \text{“}text\text{”} & .02 & .001 & \dots & 0.001 \\ \text{“}data\text{”} & .001 & .02 & \dots & 0.001 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{“}analysis\text{”} & .01 & .01 & \dots & 0.0005 \end{array}\right]$$

Topic models

For Latent Dirichlet Allocation...

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

Topic models

For Latent Dirichlet Allocation...

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

Topic models

For Latent Dirichlet Allocation...

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

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

Topic models

For Latent Dirichlet Allocation...

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

$$\begin{aligned} \boldsymbol{\theta}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\ \boldsymbol{z}_{im} | \boldsymbol{\theta}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}_i) \end{aligned}$$

Topic models

For Latent Dirichlet Allocation...

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

$$
\begin{aligned}
\boldsymbol{\theta}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{z}_{im} | \boldsymbol{\theta}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}_i) \\
w_{im} | \boldsymbol{\beta}_k, z_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\beta}_k)
\end{aligned}
$$

Topic models

For Latent Dirichlet Allocation...

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

$$\begin{aligned}
\boldsymbol{\beta}_k &\sim \text{Dirichlet}(\mathbf{1}) \\
\\
\boldsymbol{\theta}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{z}_{im} | \boldsymbol{\theta}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}_i) \\
w_{im} | \boldsymbol{\beta}_k, z_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\beta}_k)
\end{aligned}$$

Topic models

For Latent Dirichlet Allocation...

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

$$
\begin{aligned}
\boldsymbol{\beta}_k &\sim \text{Dirichlet}(\mathbf{1}) \\
\alpha_k &\sim \text{Gamma}(\alpha, \beta) \\
\boldsymbol{\theta}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{z}_{im} | \boldsymbol{\theta}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}_i) \\
w_{im} | \boldsymbol{\beta}_k, z_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\beta}_k)
\end{aligned}
$$

Topic models

For Latent Dirichlet Allocation...

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

$$
\begin{aligned}
\boldsymbol{\beta}_k &\sim \text{Dirichlet}(\mathbf{1}) \\
\alpha_k &\sim \text{Gamma}(\alpha, \beta) \\
\boldsymbol{\theta}_i | \boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \\
\boldsymbol{z}_{im} | \boldsymbol{\theta}_i &\sim \text{Multinomial}(1, \boldsymbol{\theta}_i) \\
w_{im} | \boldsymbol{\beta}_k, z_{imk} = 1 &\sim \text{Multinomial}(1, \boldsymbol{\beta}_k)
\end{aligned}
$$

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/` > "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.

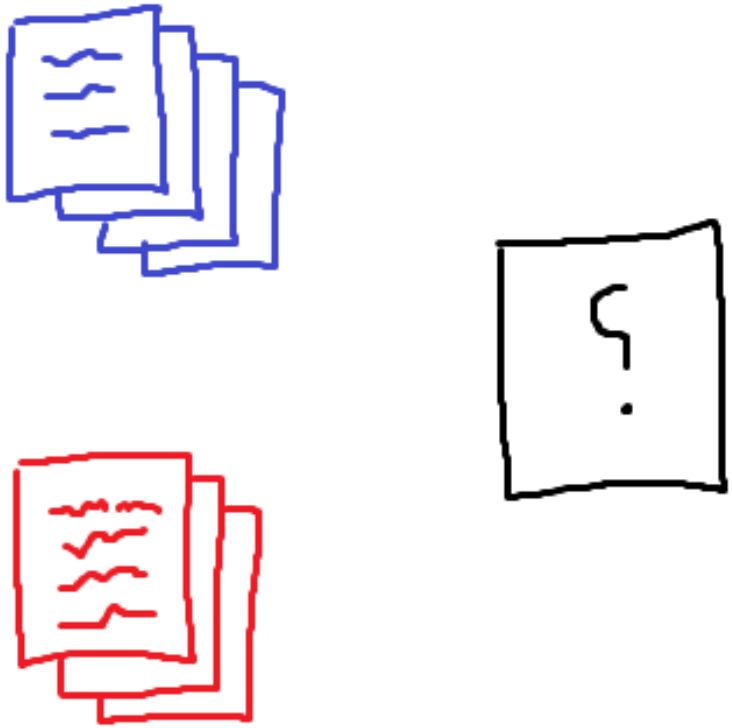
Classification

Goal: Predict document type from text.

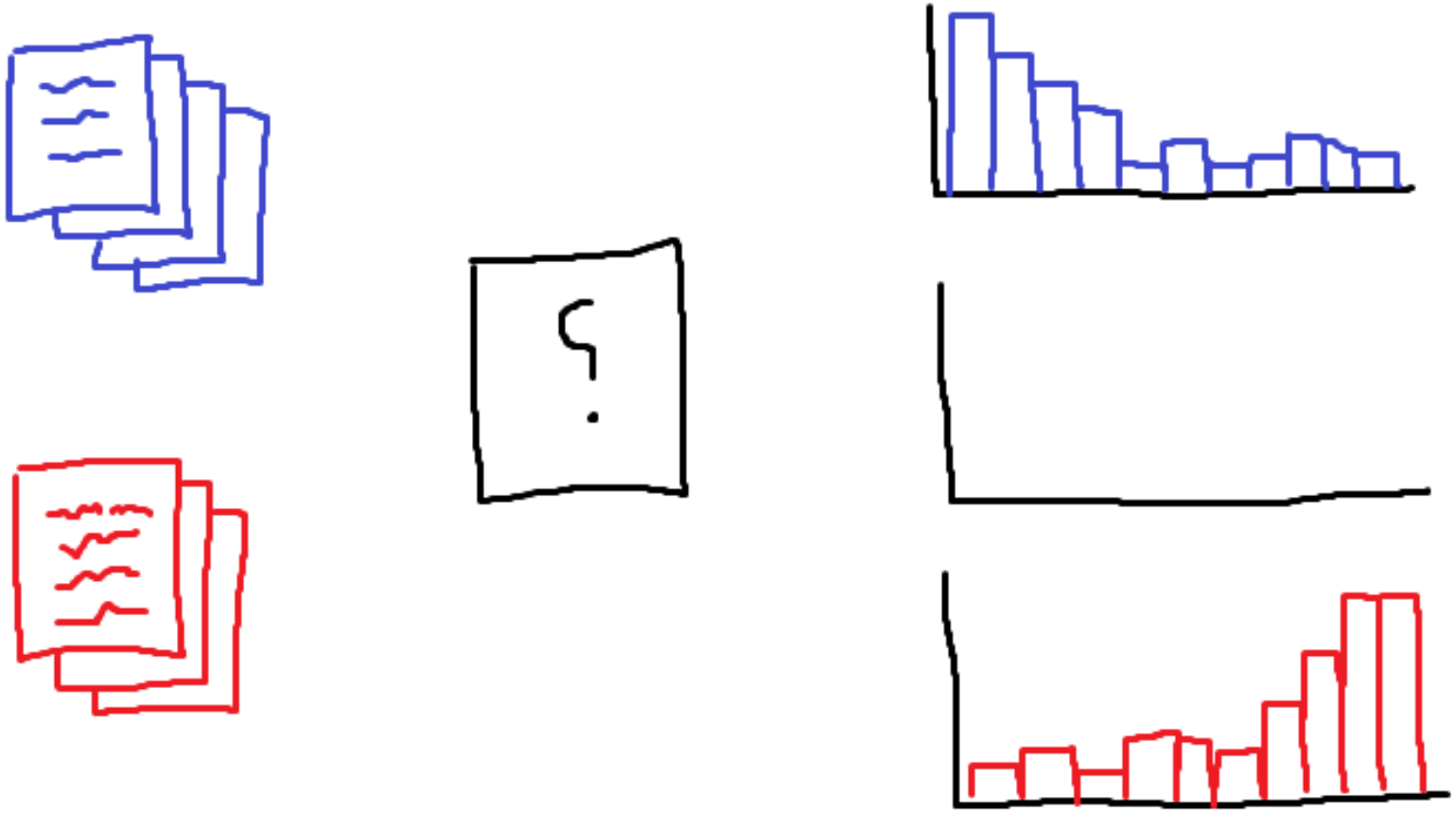
Good: You know what you are getting. Faster than hand-coding everything.

Bad: Time consuming to label documents.

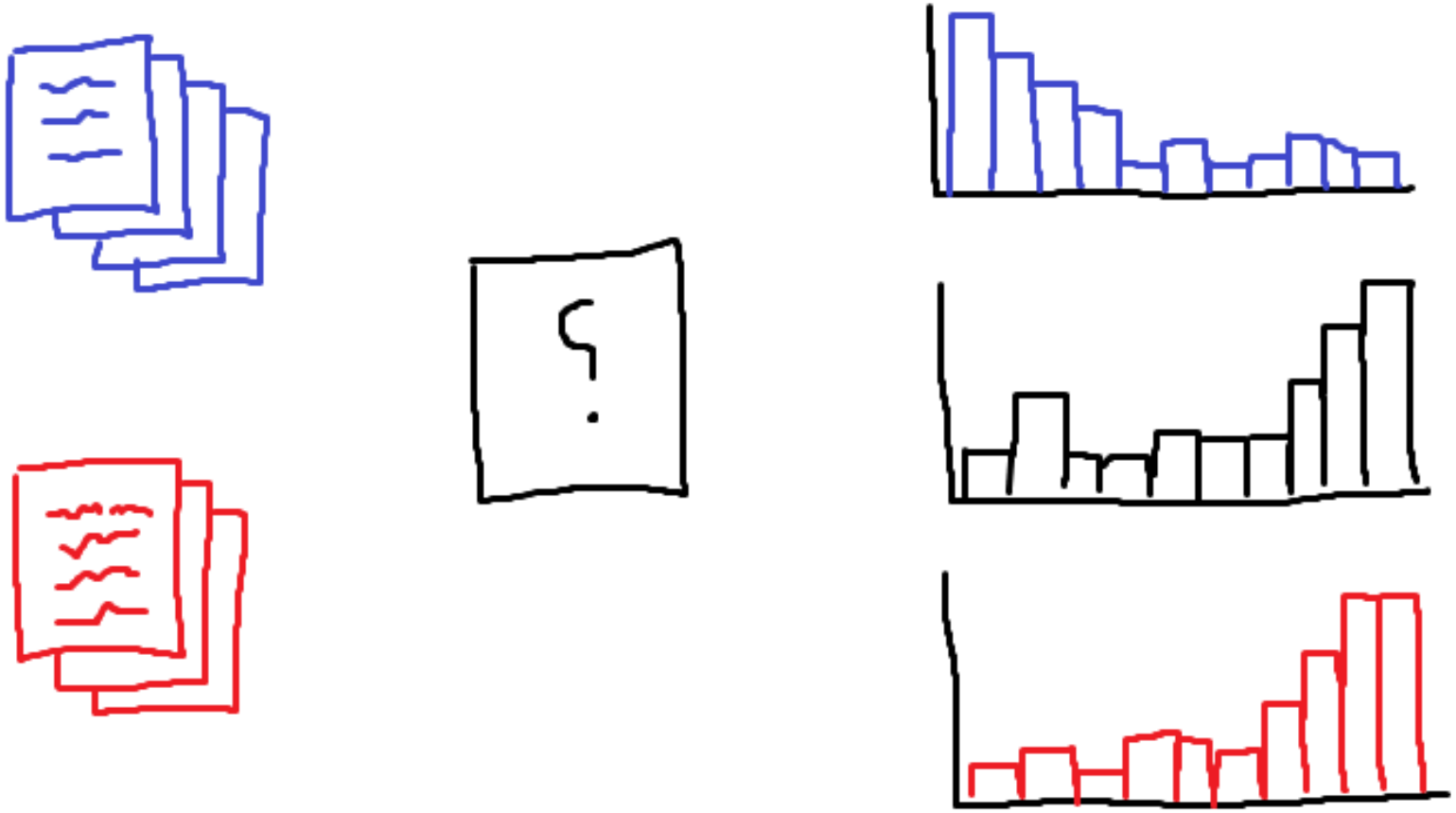
Classification



Classification

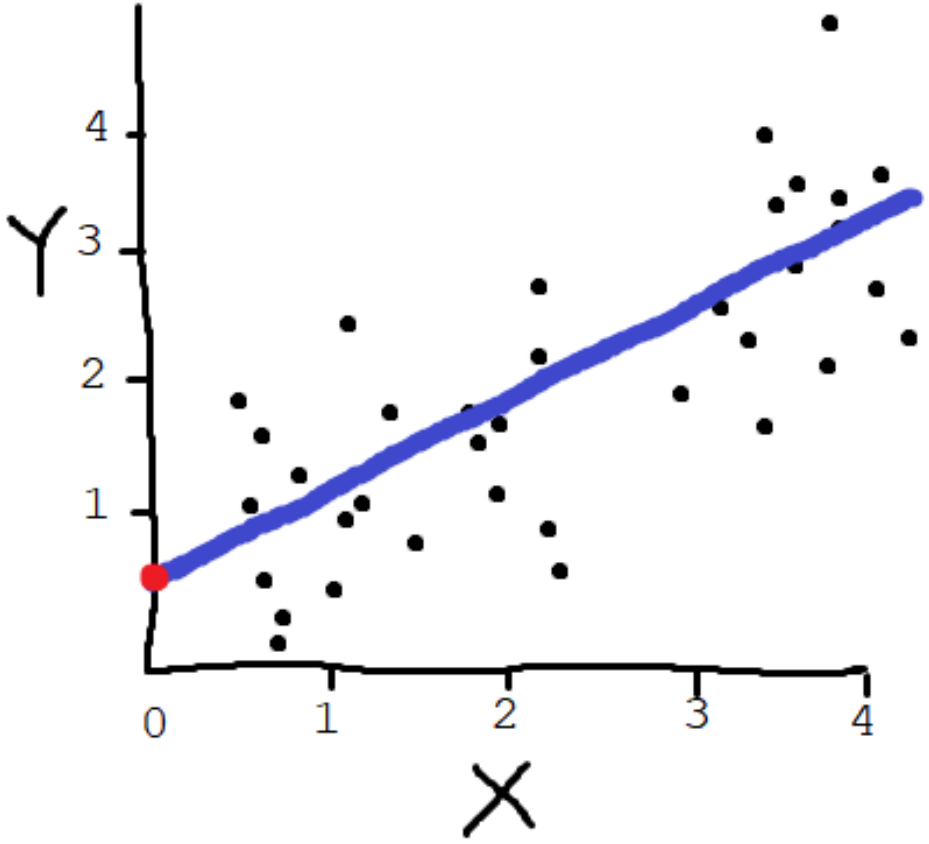


Classification

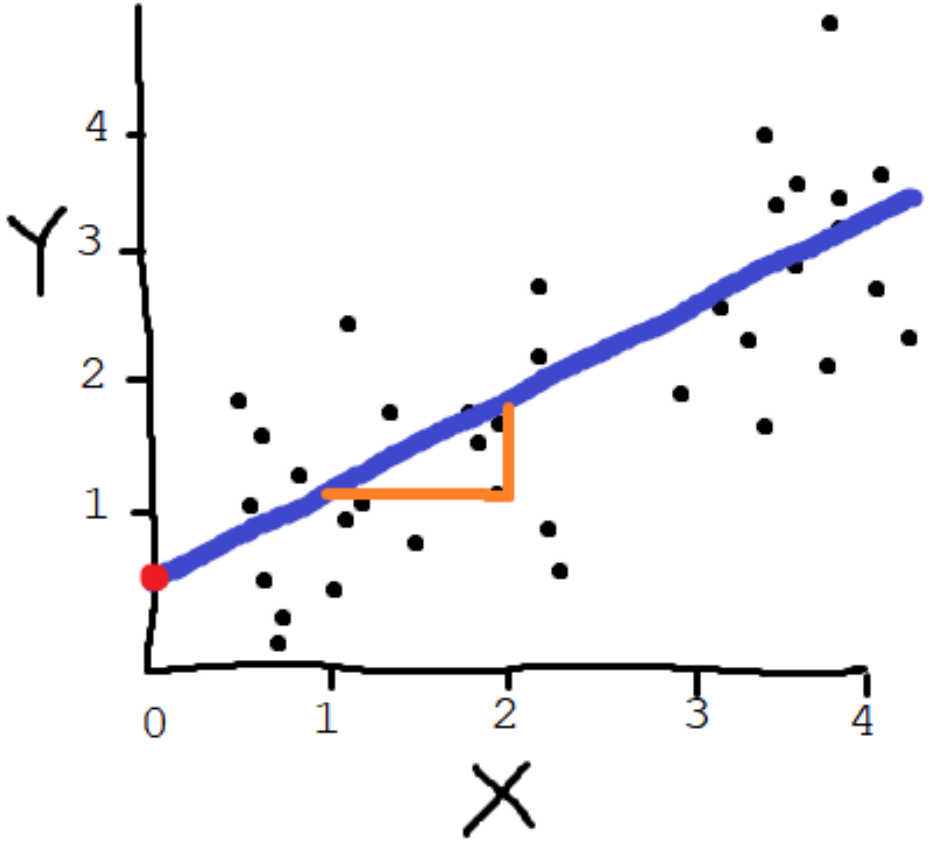


Classification

Classification

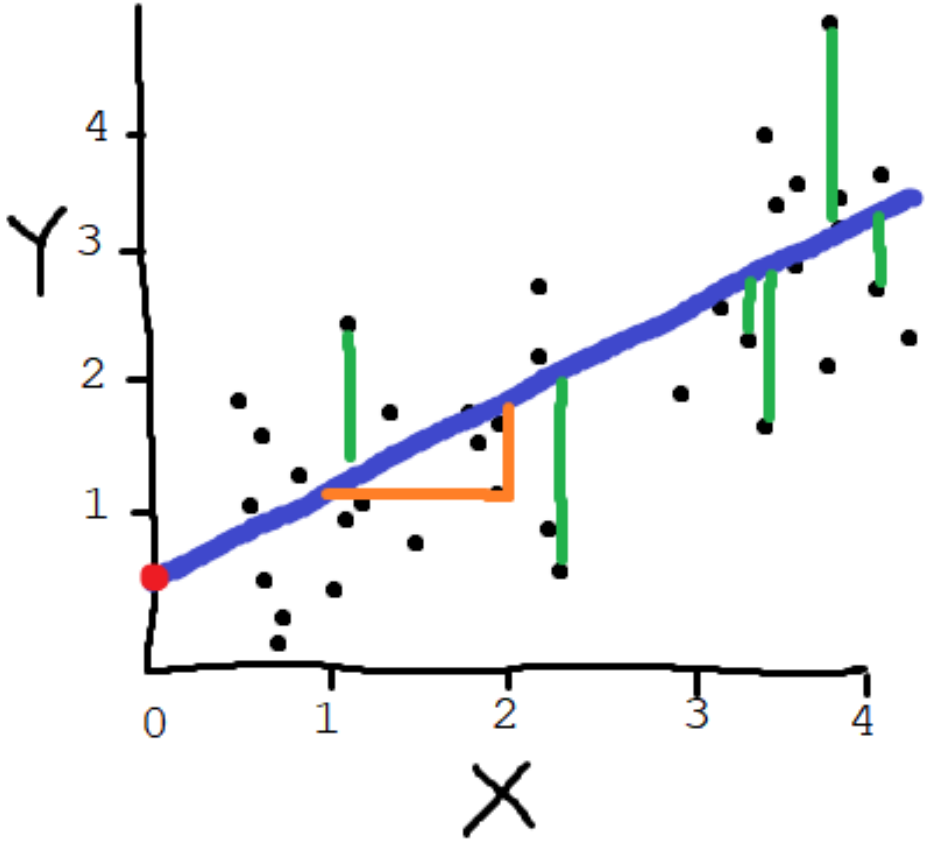


Classification



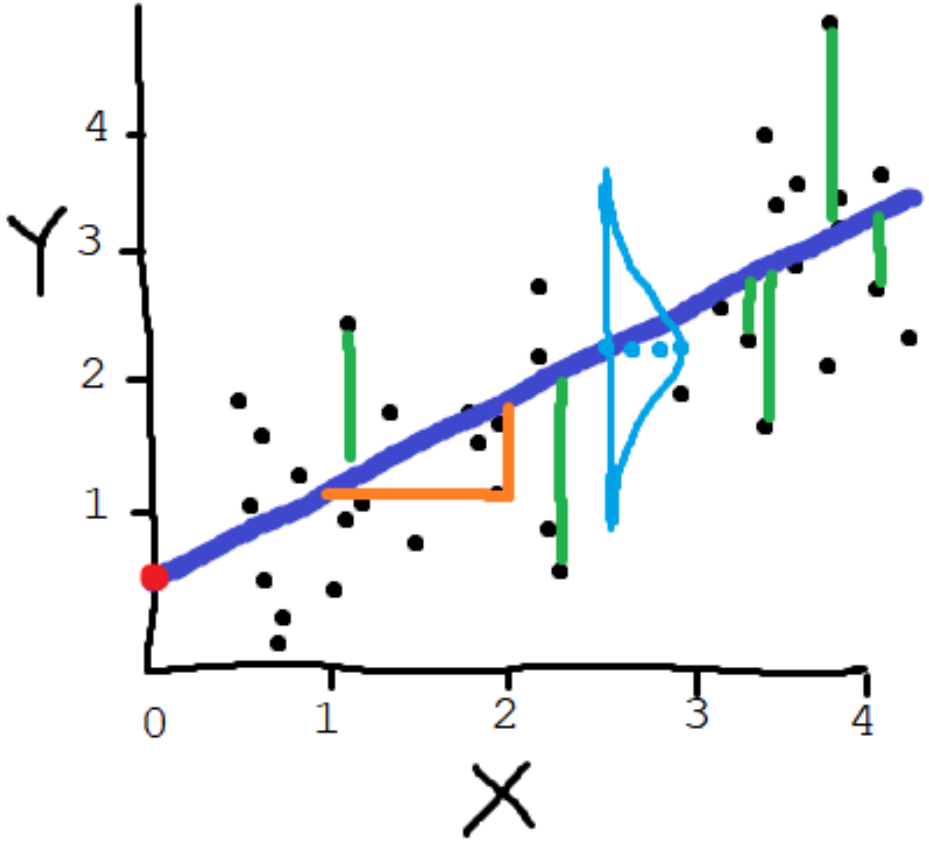
$$y = \alpha + \beta x$$

Classification



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

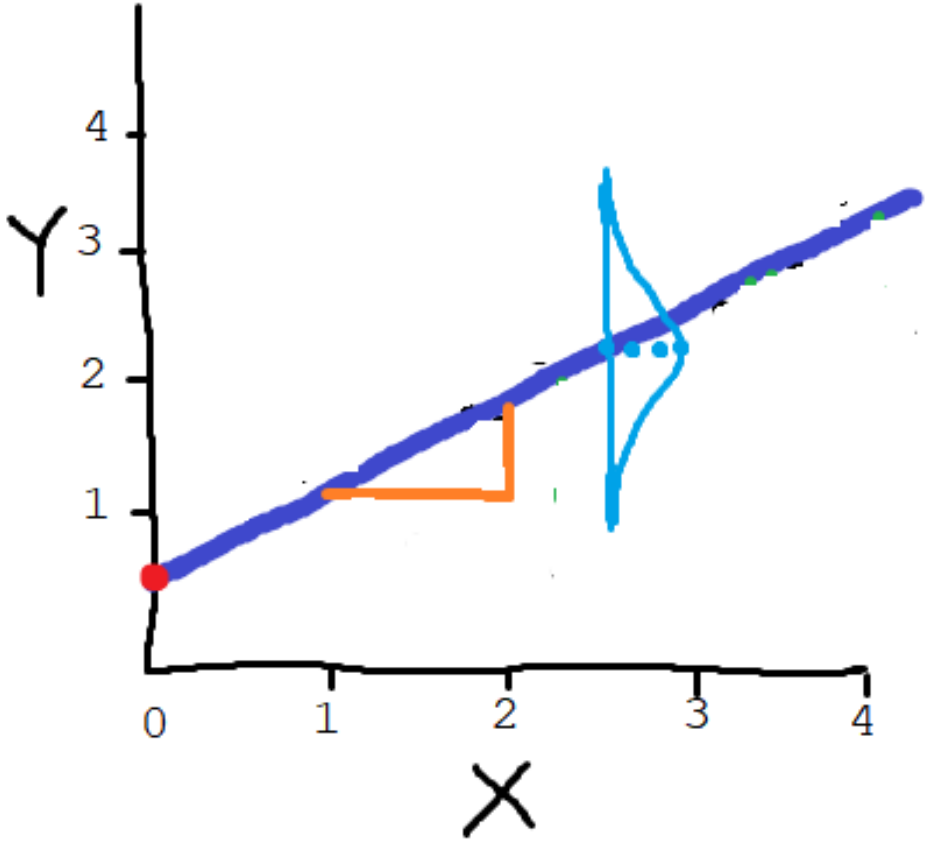
Classification



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

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

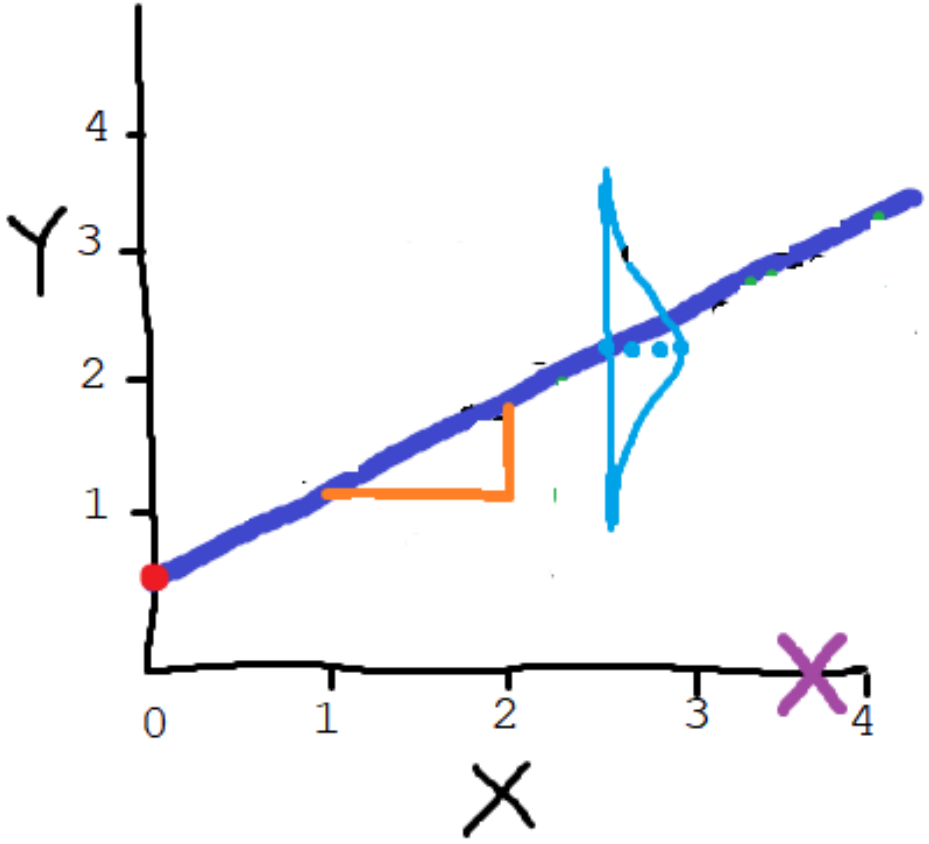
Classification



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

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

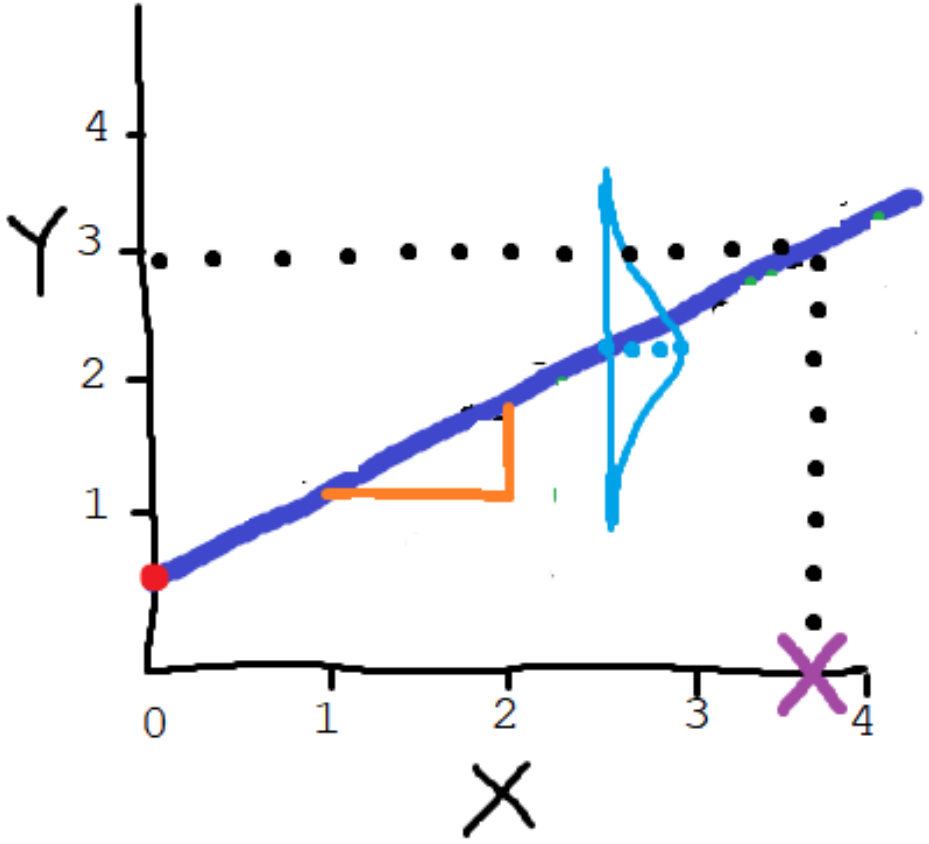
Classification



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

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

Classification

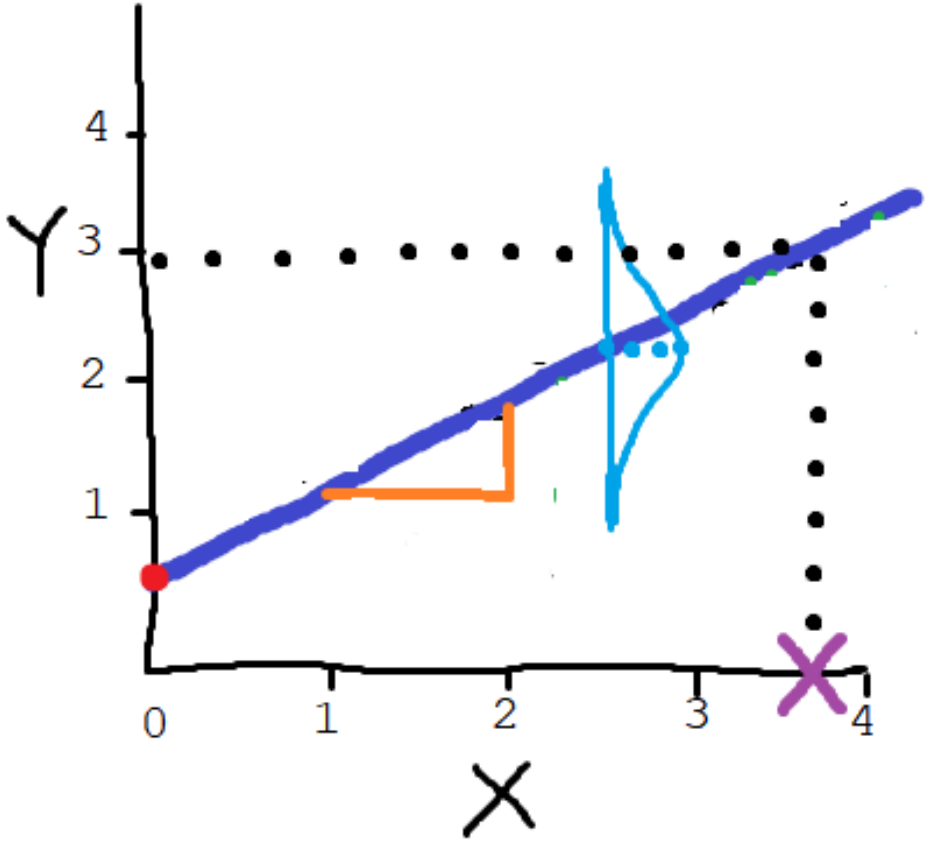


$$y = \alpha + \beta x + \epsilon$$

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

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

Classification



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

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

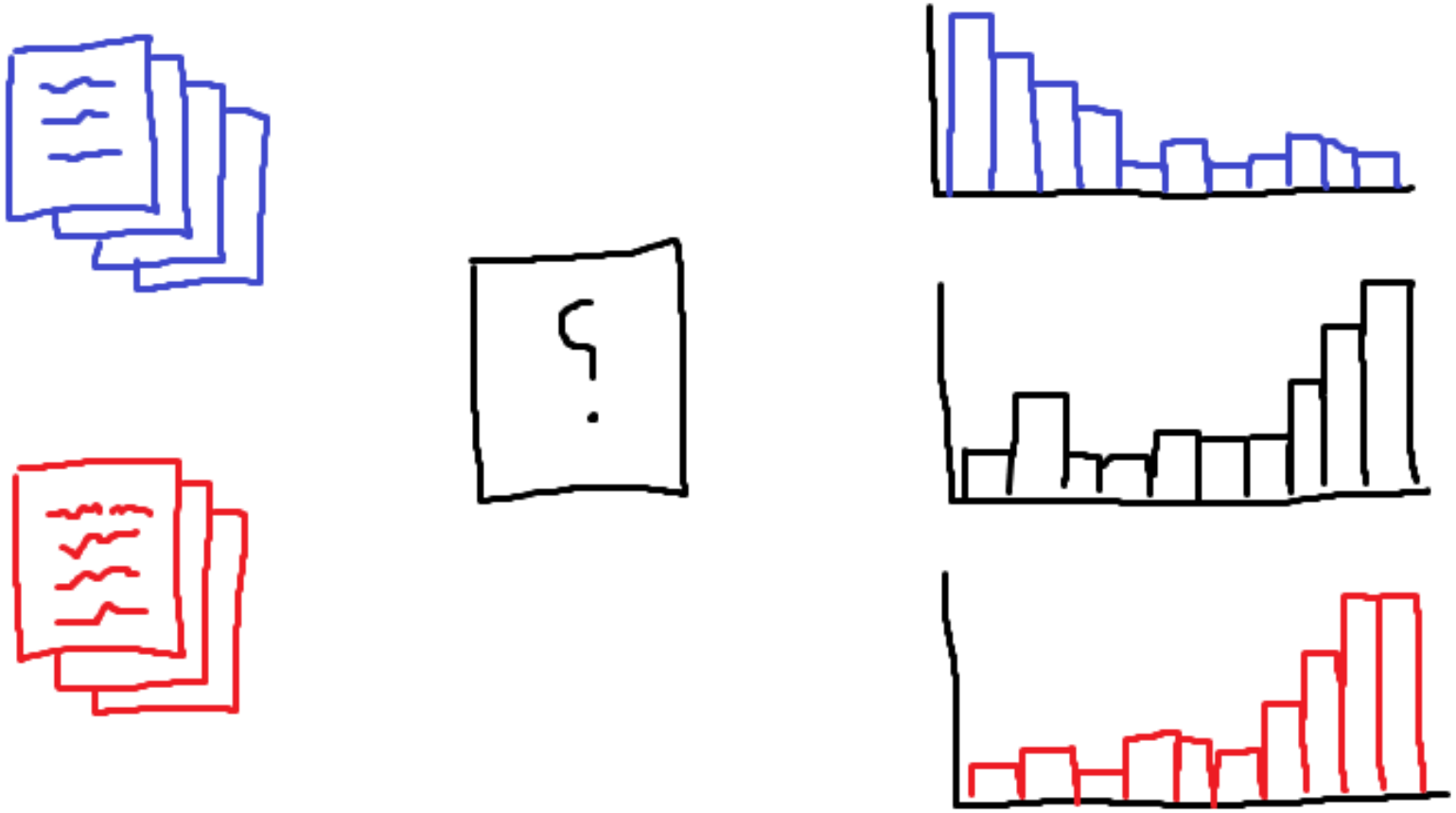
$$\hat{y} = \hat{\alpha} + \hat{\beta} x$$

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:

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

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

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/` > "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:

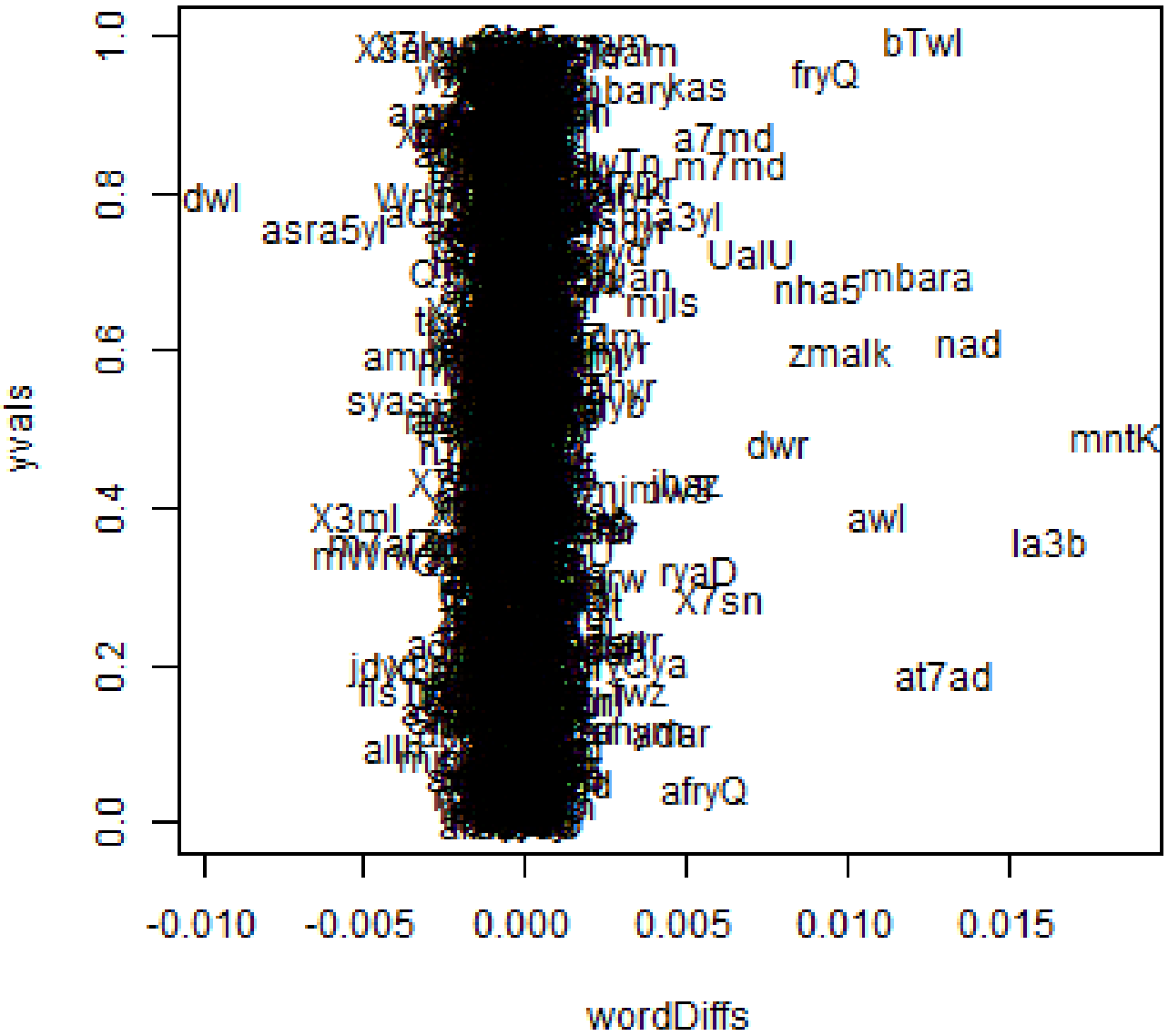
- Every graph should have a point
- Clear labels
- Every element should serve a purpose
- Avoid redundant information
- Avoid misleading

Visualization

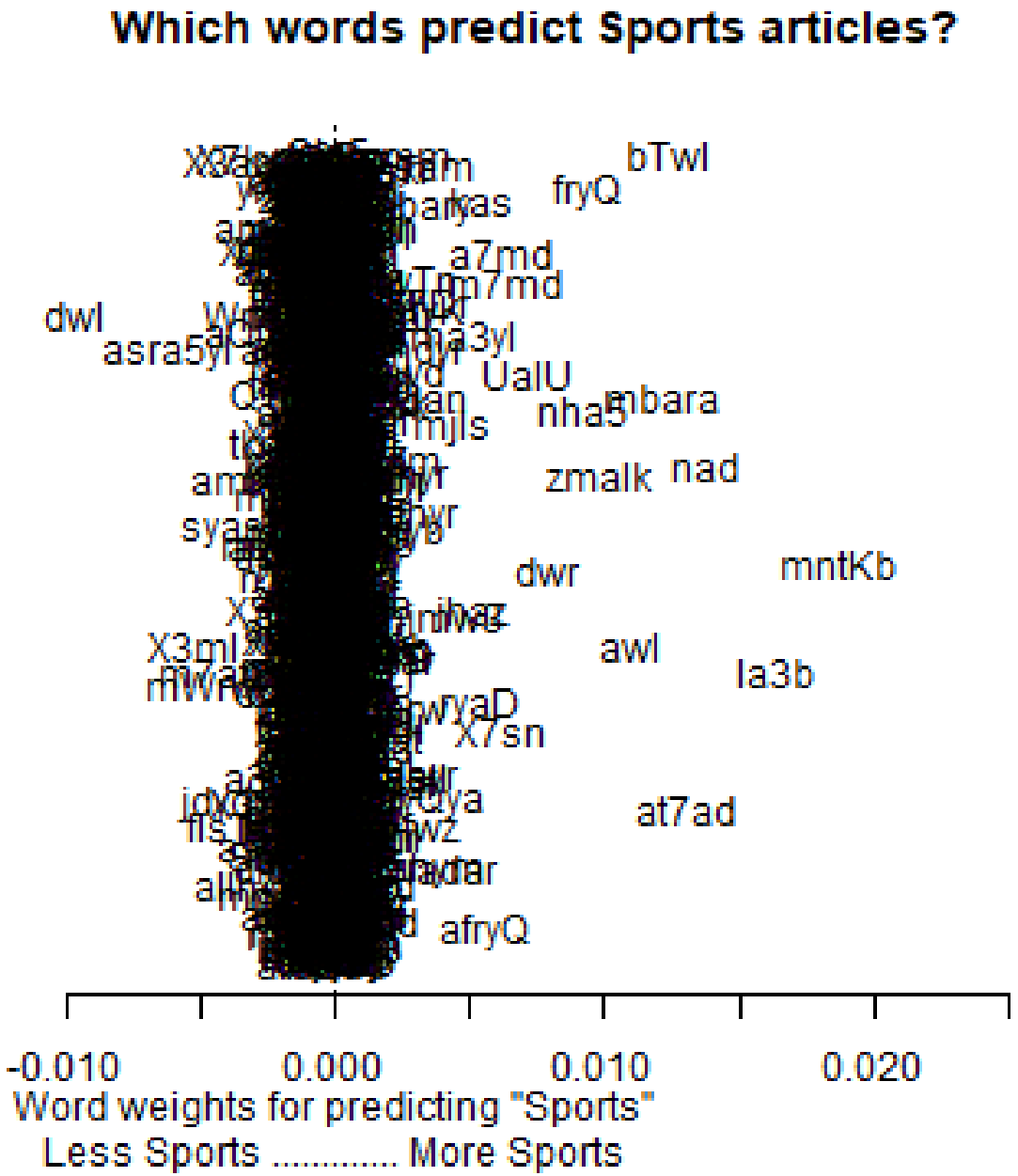
R Exercise. Code: `http://www.mit.edu/~rnielsen/arabicTextWorkshop.zip`

Slides: `http://www.mit.edu/~rnielsen/arabic_text_slides.pdf`
`http://www.mit.edu/~rnielsen/` > "Helpful Stuff" (at the bottom)

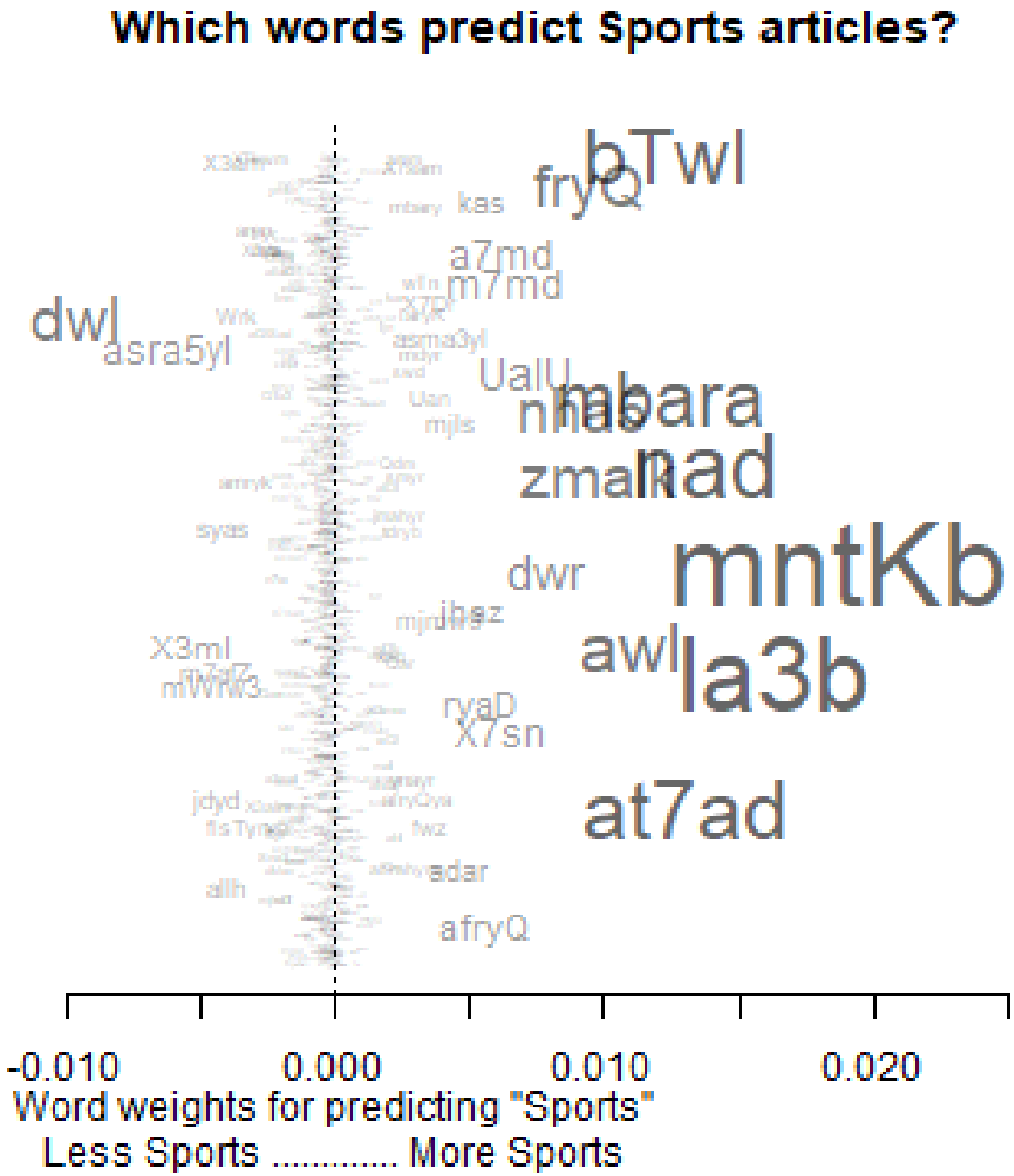
Visualization: Naive Bayes Classifier



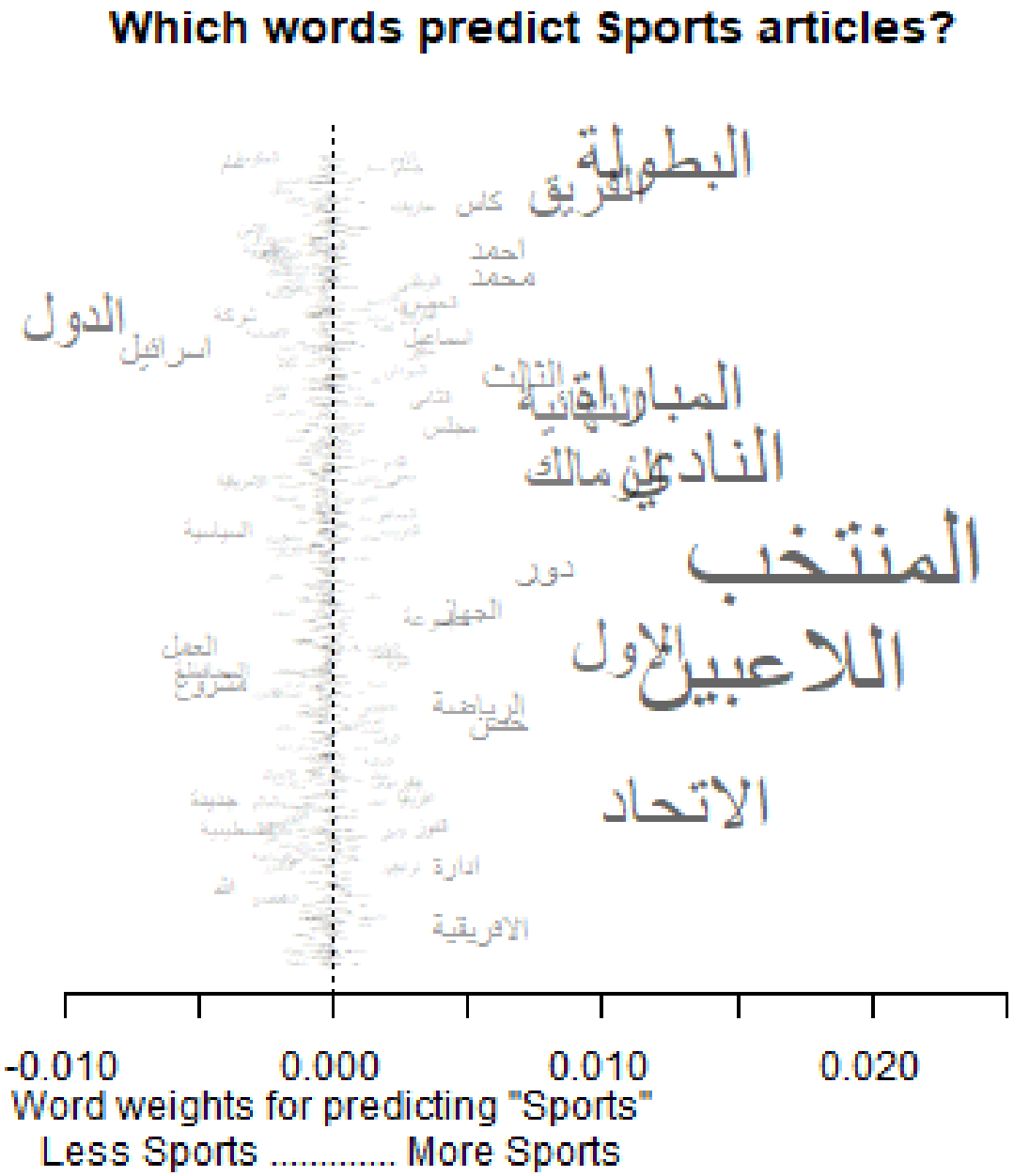
Visualization: Naive Bayes Classifier



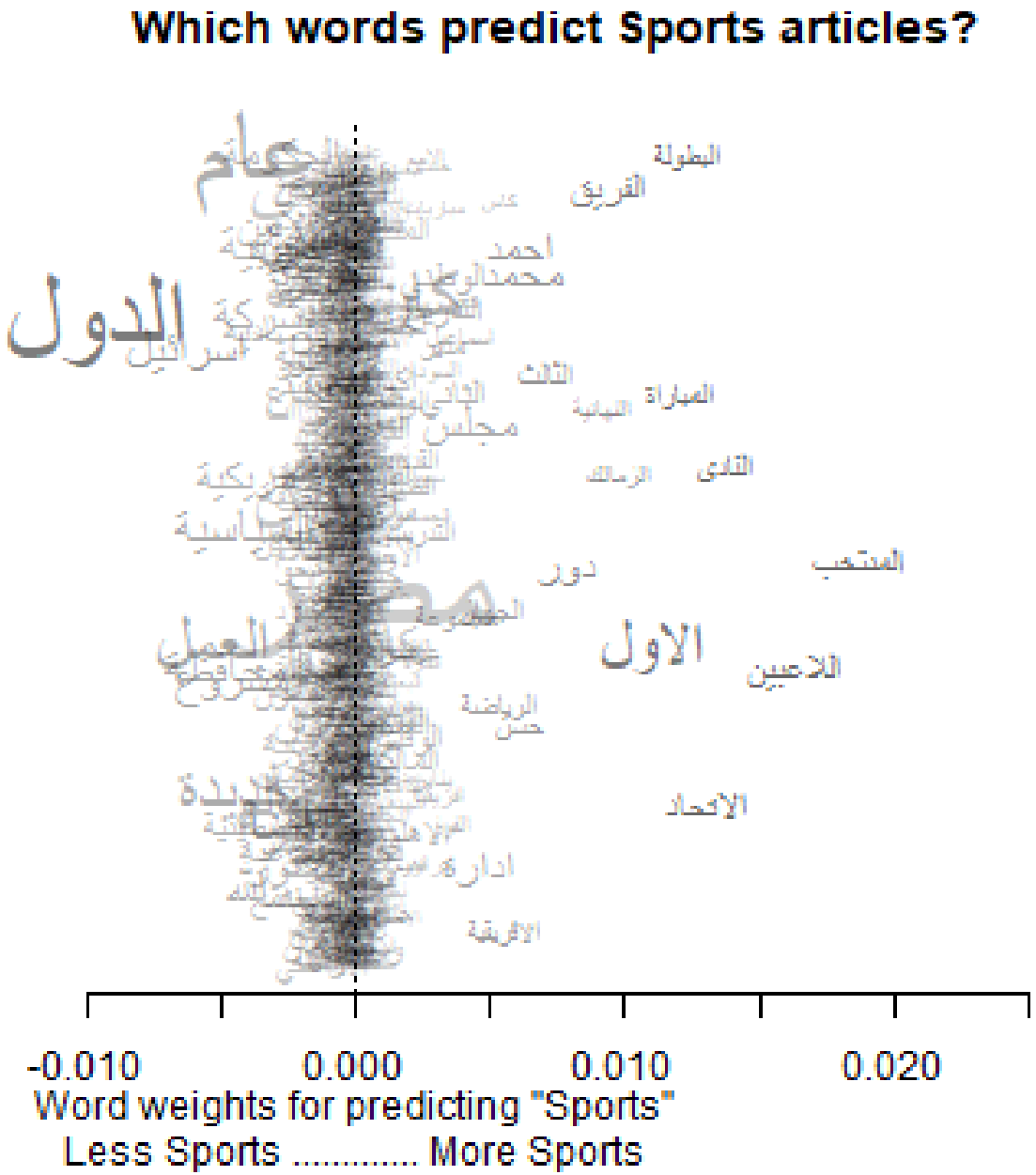
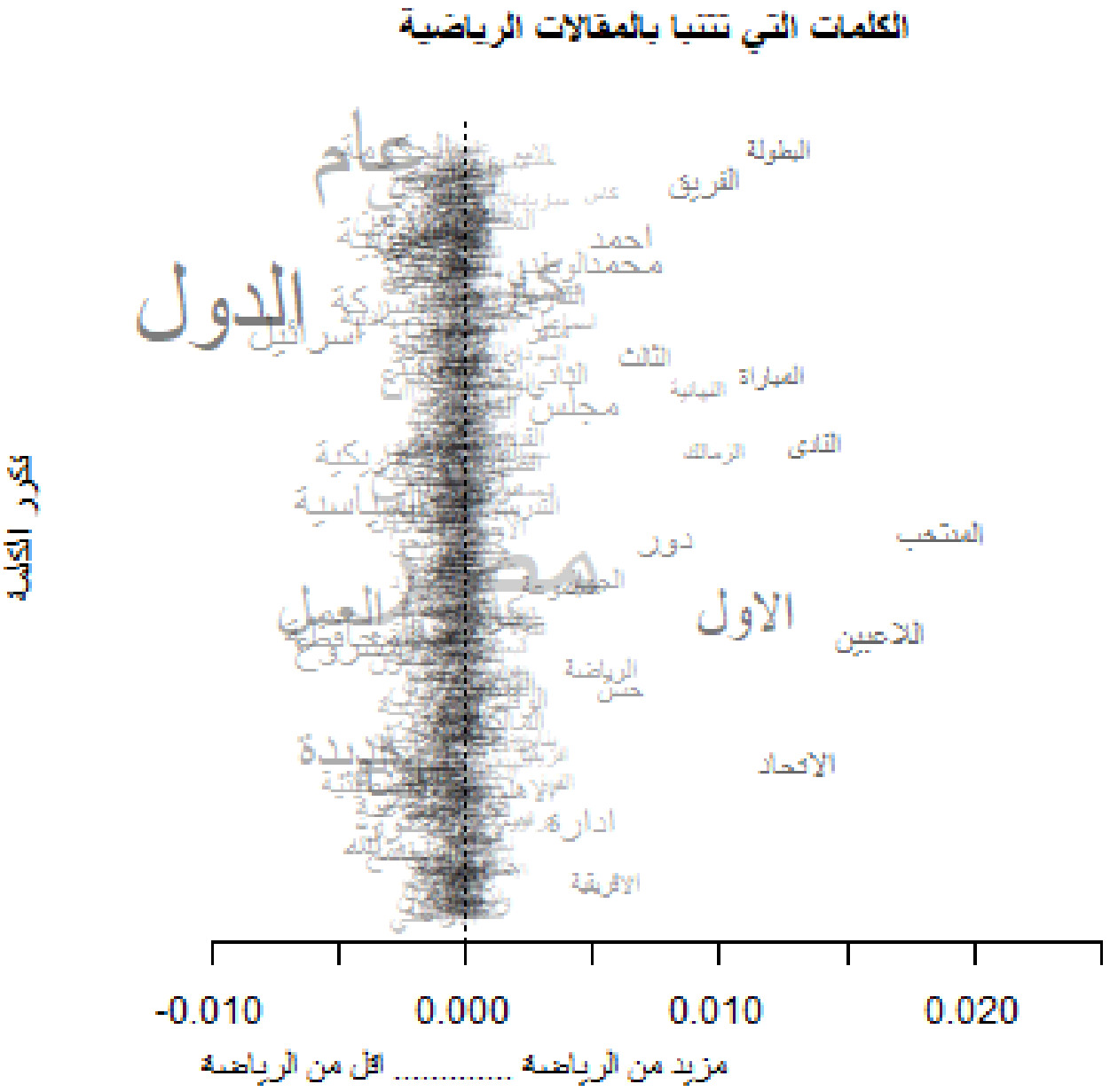
Visualization: Naive Bayes Classifier



Visualization: Naive Bayes Classifier



Visualization: Naive Bayes Classifier



Visualization: Naive Bayes Classifier