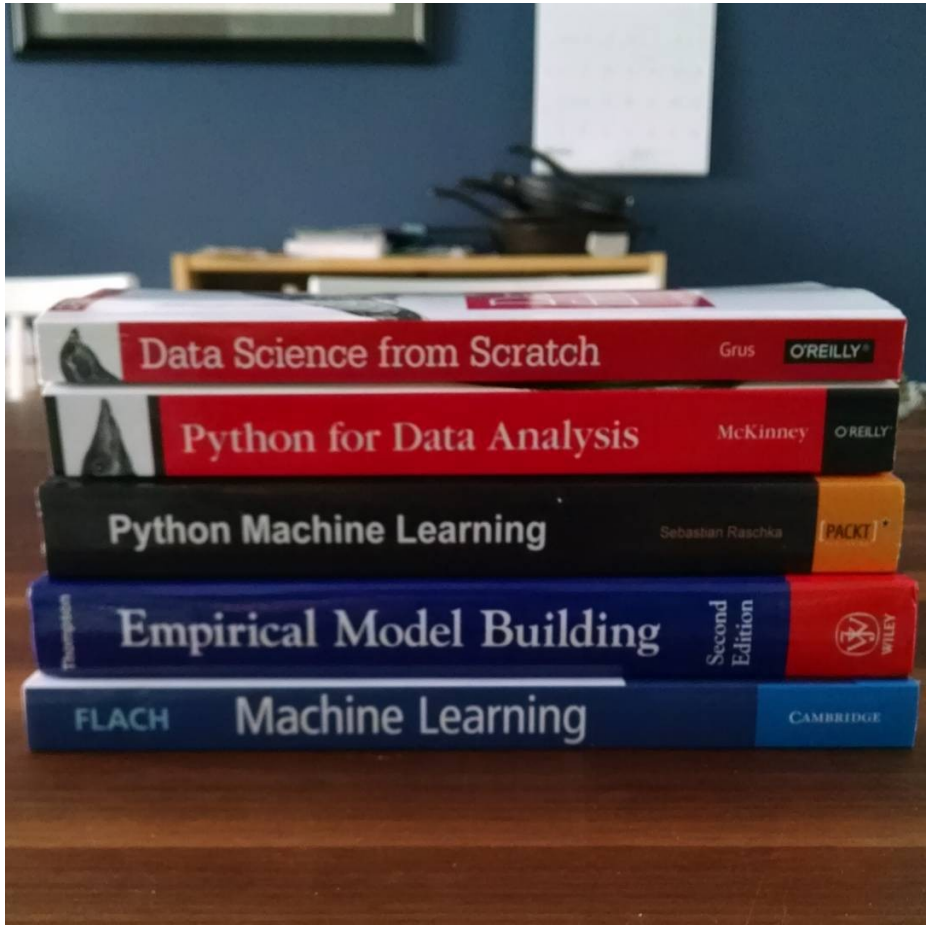
Keyword Mining without Stop Word Lists

PRO-DEV!


Data Science from Scratch
Grus
O'REILLY
Python for Data Analysis
McKinney
O'REILLY
Python Machine Learning
Sebastian Raschka
PACKT
Empirical Model Building
Second Edition
WILEY
FLACH
Machine Learning
CAMBRIDGE


coursera

Document Similarity

If you like this article about Messi, you might be interested in this one about Pelé

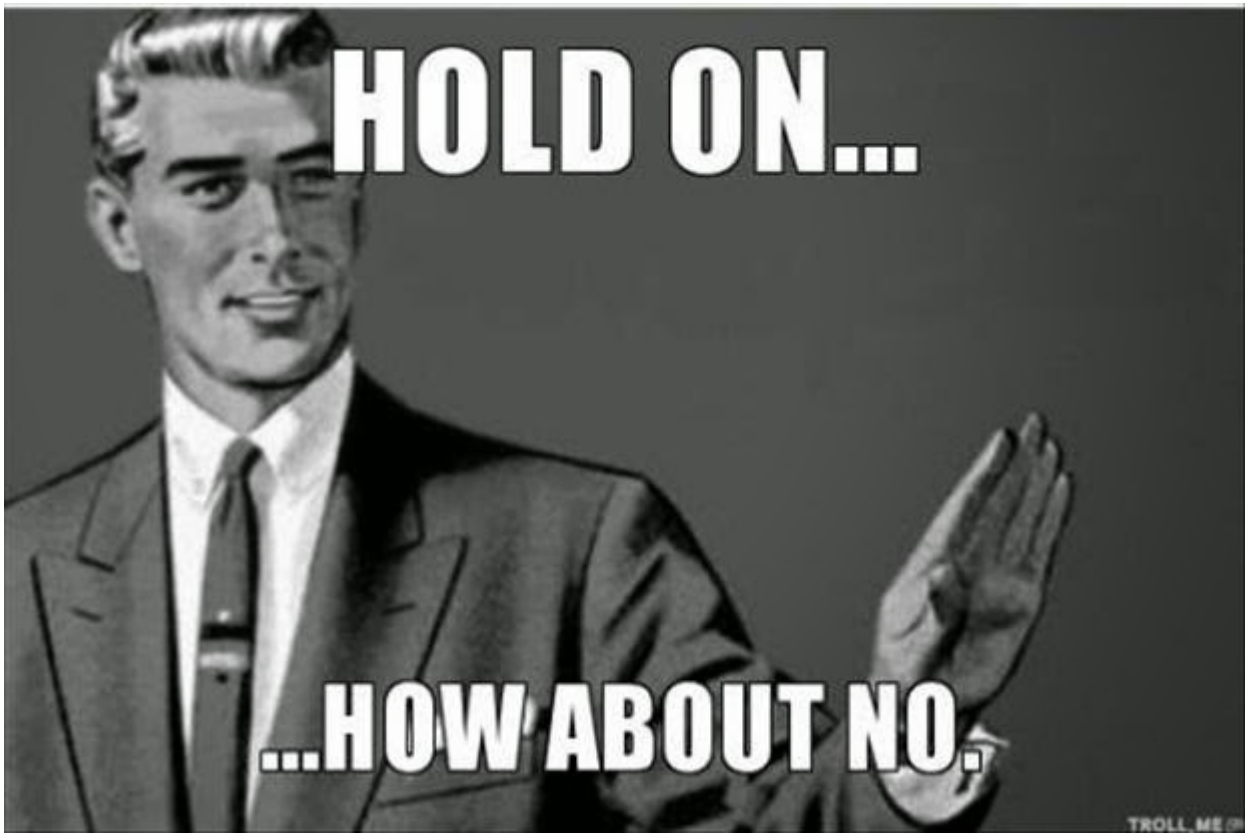
Document Similarity

- Soccer
- Football
- South
- America
- World
- Cup

Document Similarity

- Soccer
- Football
- South
- America
- World
- Cup

- the
- and
- Wikipedia
- Privacy
- Policy
- Terms


HOLD ON...
...HOW ABOUT NO.
TROLL ME

What is a keyword in a document?

A term that appears frequently within a particular document, but not so frequently in other documents

tf-idf

Term frequency-Inverse document frequency

https://en.wikipedia.org/wiki/Tf%E2%80%93idf

Term Frequency

How often a term appears within a document indicates how “important” it is to that document

{ 'the': 789, 'bocoup': 456, 'loft': 123 }

In practice, we normalize this, because longer documents have more occurrences of words

Inverse Document Frequency

How “rare” a term is within the corpus

math.log(number_of_docs / docs_with_term)

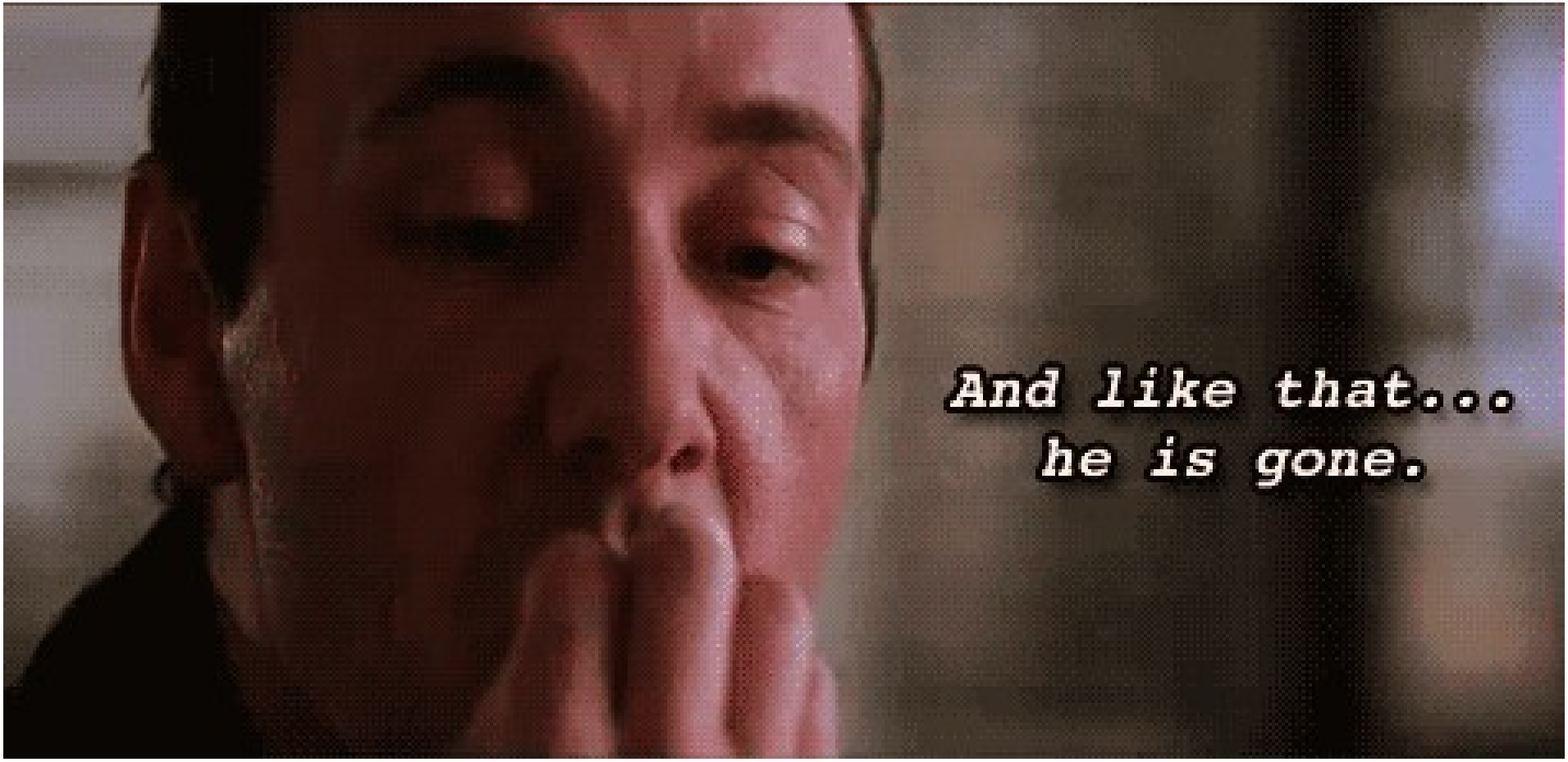
{ 'the': 0, 'bocoup': 0.12, 'loft': 0.84 }

tf-idf

tf = { 'the': 789, 'bocoup': 456, 'loft': 123 }
idf = { 'the': 0, 'bocoup': 0.12, 'loft': 0.84 }

tf-idf = { 'the': 0, 'bocoup': 54.72, 'loft': 103.32 }

Stop Words



Bocoup.com

bocoup.com/about/bocouper/jasmin-jata

- ceramic
- rose
- kelleher
- cephalopod
- reveling
- biology
- scientific
- outlook
- unknown
- overlords

bocoup.com/weblog/category/johnny-five

- milliamps
- paired
- measurement
- programmed
- electronic
- lipo
- easilydiscoverable
- motor
- pasting
- precious

bocoup.com/weblog/sponsoring-diversity-scholarships-for-braziljs

- scholarship
- applicants
- welcomed
- eligible
- heartfelt
- evaluators
- vented
- devoted
- perspectives
- accomodation

Next Steps and Applications

- Recommendations (Show related blogposts without tags)
- What distinguishes us from competitors?
- SEO/search terms

GitHub

https://github.com/bobholt/ml