BIG DATA ANALYSIS & VISUALISATION

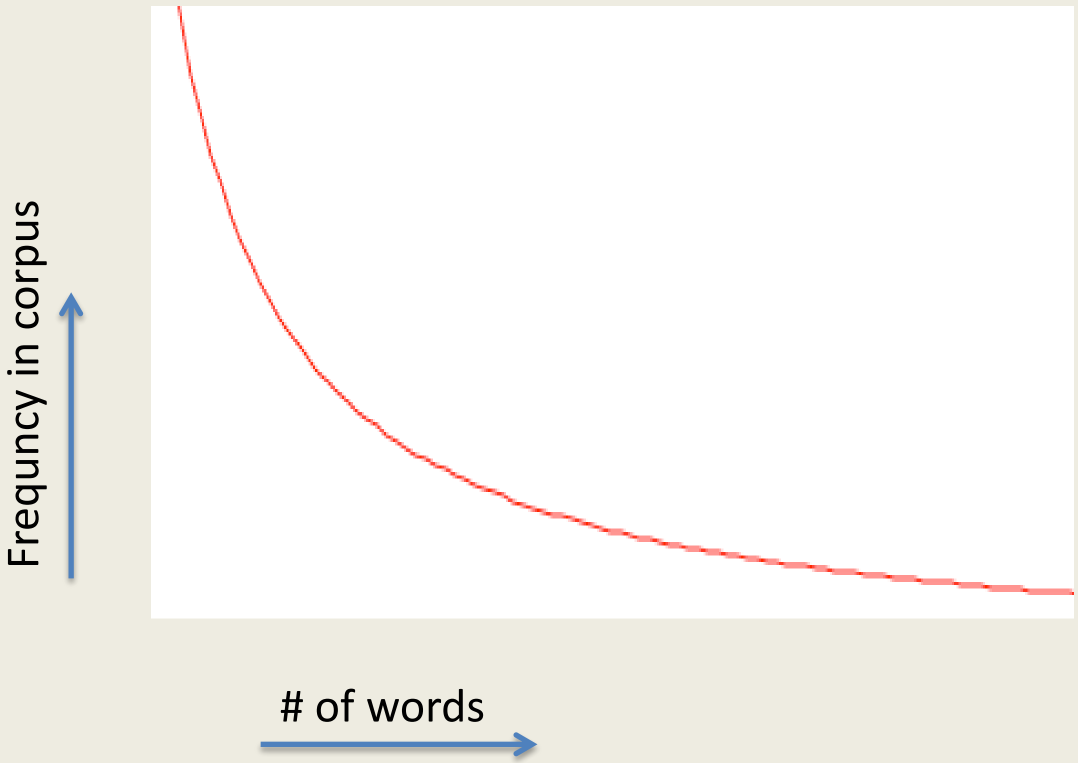
Text Mining / Text Analytics

EGS course
in collaboration with Erasmus Studio

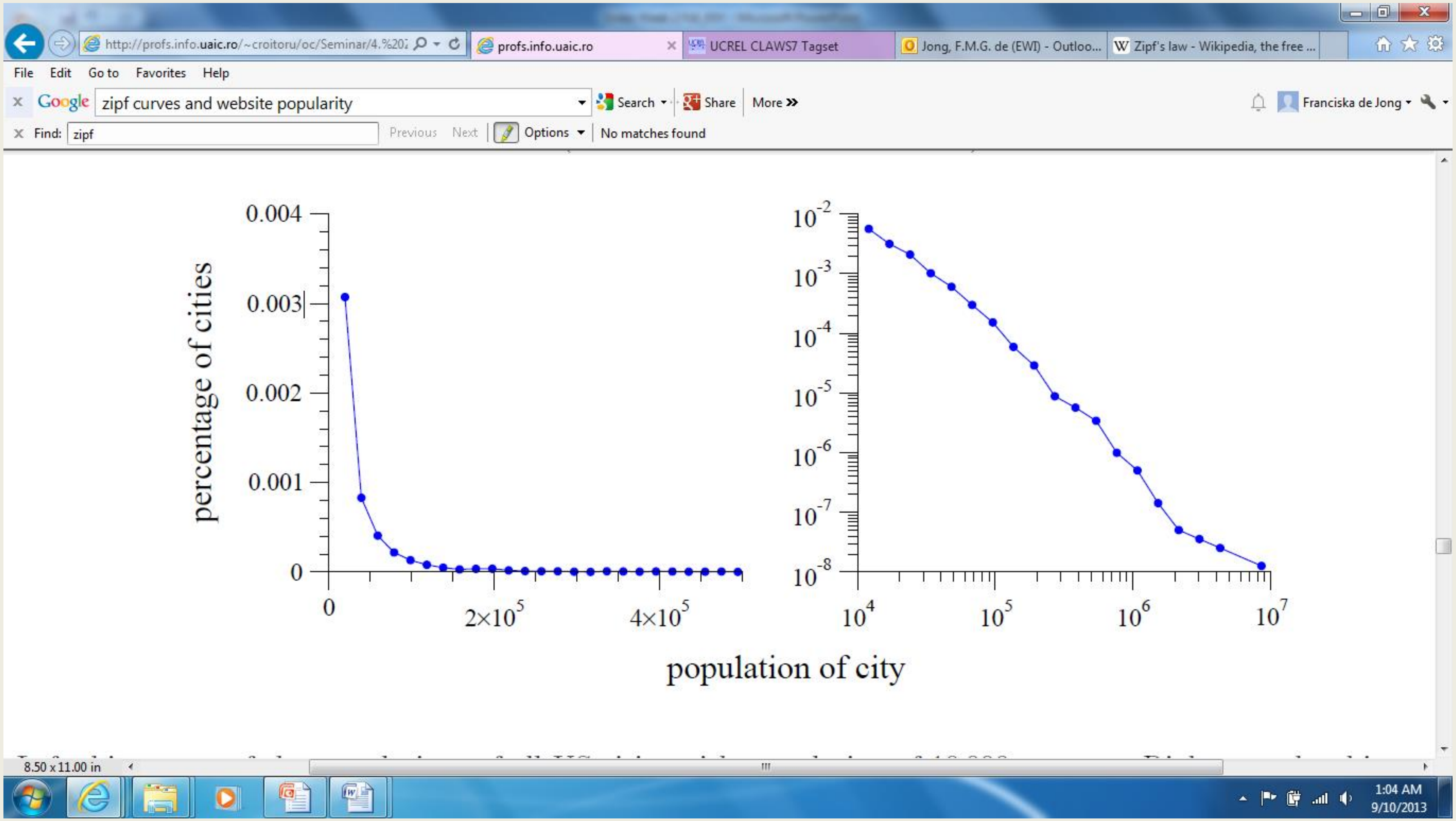
Zipf’s Law (1)

- Words follow a Zipfian distribution
  - A small number of words occurring very frequently
  - A large number of words occurring rarely
- In math-style: *a word’s frequency is approximately inversely proportional to its rank in the word distribution list.*
- The most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc.

Zipf's Law (2)



Zipf's Law (3)



http://profs.info.uaic.ro/~croitoru/oc/Seminar/4.%20
profs.info.uaic.ro
UCREL CLAWS7 Tagset
Jong, F.M.G. de (EWI) - Outloo...
Zipf's law - Wikipedia, the free ...
File Edit Go to Favorites Help
zipf curves and website popularity
Search Share More
Franciska de Jong
Find: zipf
Previous Next Options No matches found
percentage of cities
0.004
0.003
0.002
0.001
0
2×10^5
4×10^5
10^-2
10^-3
10^-4
10^-5
10^-6
10^-7
10^-8
10^4
10^5
10^6
10^7
population of city
8.50 x 11.00 in
1:04 AM
9/10/2013

Text mining: counting with words

- a digital text collection: offline, protected access (deep web, dark web) or open access
- optional step: preprocessing (data cleaning, spelling harmonisation, etc.)
- simple statistics allow simple visualizations: word frequencies (Wordles), variation over time (Ngram viewer)

Example: Ngram viewer (Google Books)

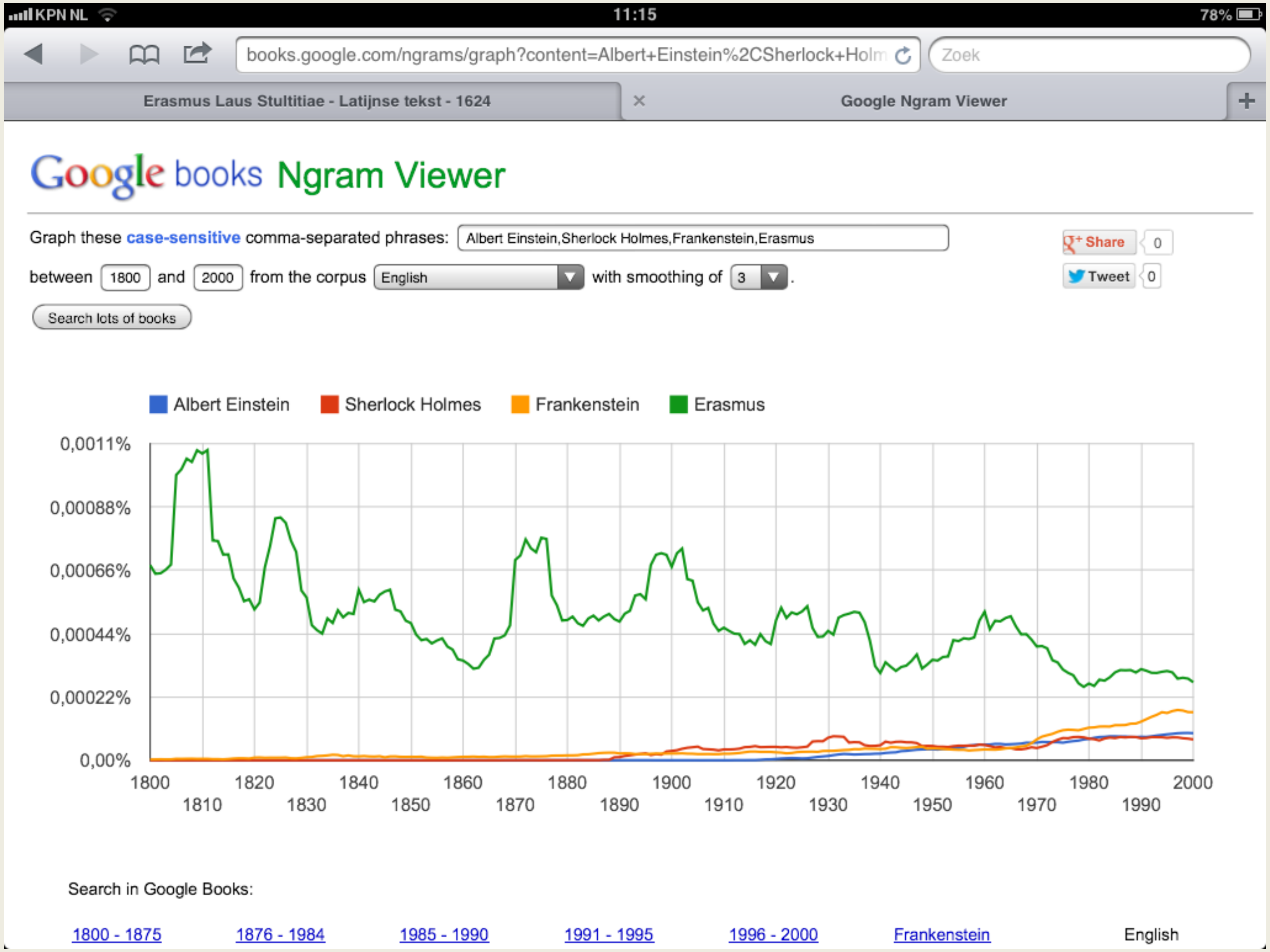
- http://books.google.com/ngrams/info
  with suggestions for how to compose very refined queries on the GB corpus (highly recommended)

- Ngram: a contiguous sequence of *n* items (units) from a given piece of text or speech. Items can phonemes, characters, syllables or words
- Jargon: an n-gram of size 1 is called *unigram*, of size 2 *bigram*, of size 3 *trigram*.

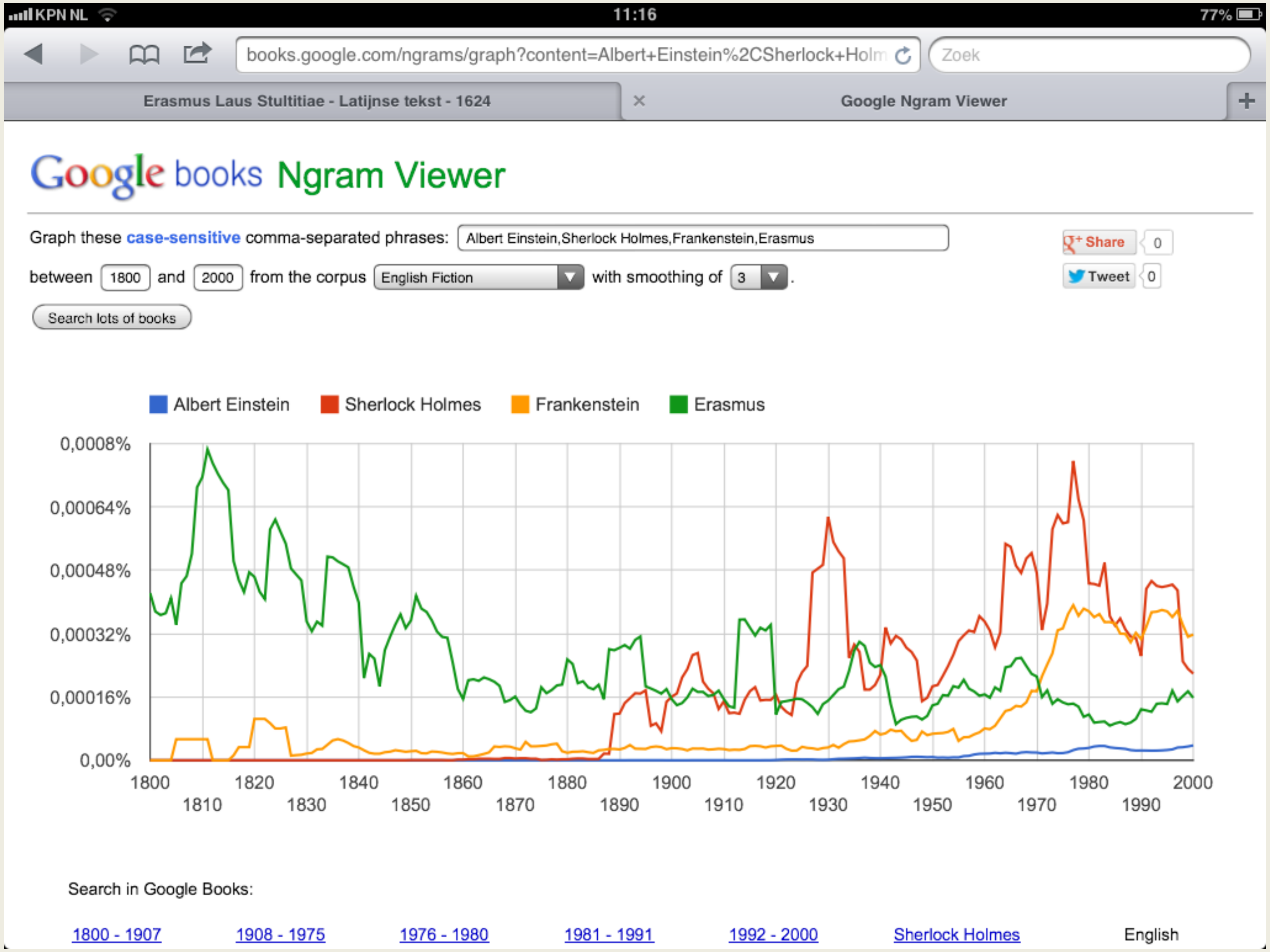
erasmus studio

Ngram viewer (1)

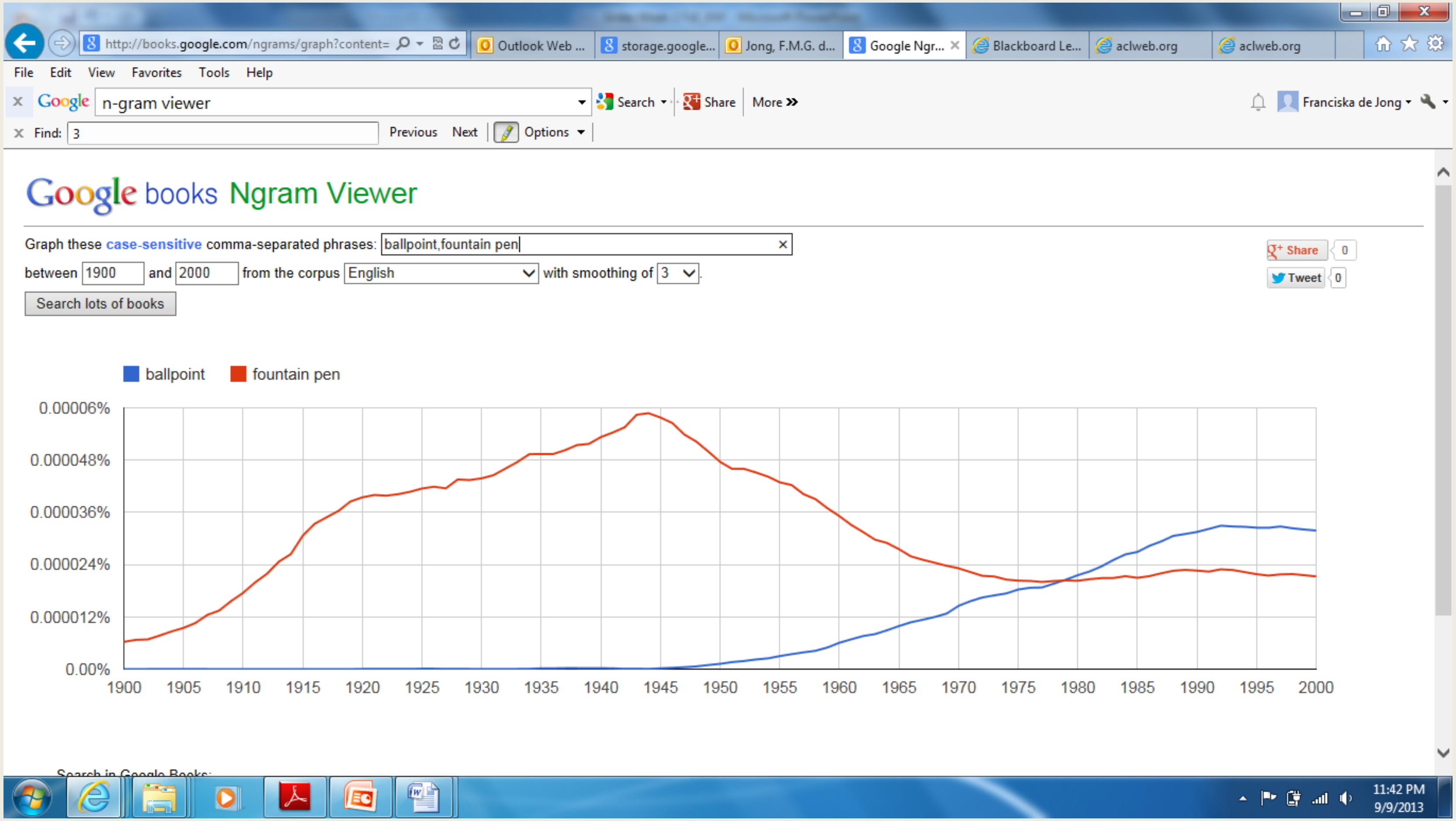
Ngram viewer (2)



Ngram viewer (3)



Ngram viewer (4)



Ngram viewer (5)

Some word count observations (1)

- There are 884,647 word occurrences (tokens) with 29,066 unique word forms (types), in an approximately one million word Shakespeare corpus
  - Shakespeare produced 300,000 bigram types out of 844 million possible bigrams: so, ***99.96% of the possible bigrams were never seen***
- You can quickly collect statistics on the high frequency words
- You might have to work an arbitrarily long time to get valid statistics on low frequency words

Some word count observations (2)

- In the Brown Corpus of American English text, the word *the* is the most frequently occurring word, and by itself accounts for nearly 7% of all word occurrences (69,971 out of slightly over 1 million). The second-place word *of* accounts for slightly over 3.5% of words (36,411 occurrences), followed by *and* (28,852). Only 135 vocabulary items are needed to account for half the Brown Corpus.
- The hundred most frequent words are mostly function words: articles, auxiliaries, preprositions, etc.

Six mood categories

A. Acerbi *et al (2013)*

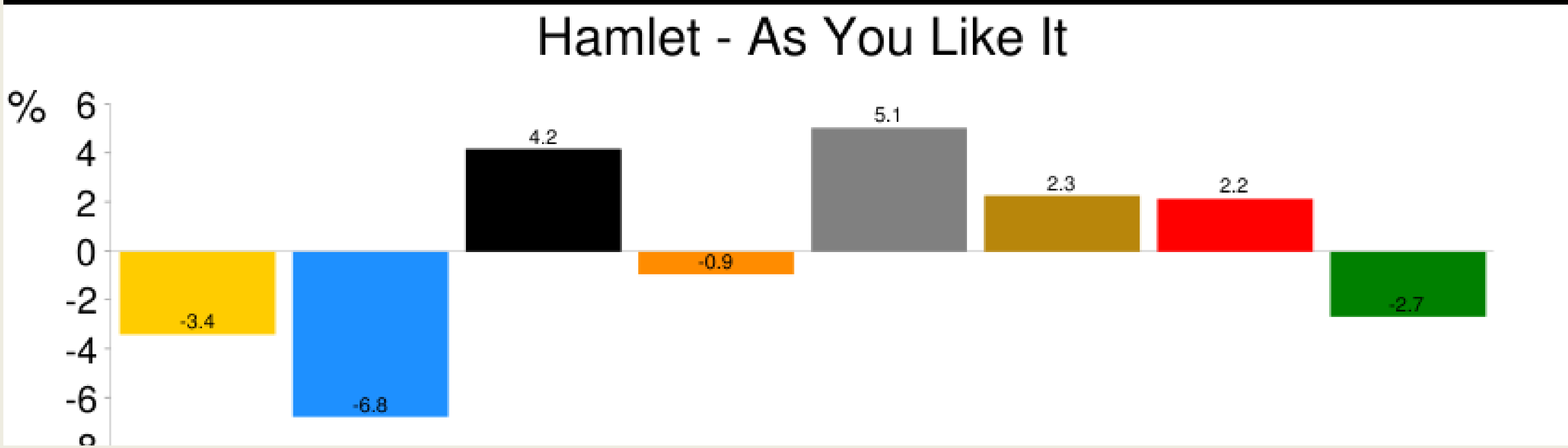
- Anger (N = 146)
- Disgust (N = 30)
- Fear (N = 92)
- Joy (N = 224)
- Sadness (N = 115)
- Surprise (N = 41)

where N stands for ….?

Mining *Hamlet* and *As you like it (1)*

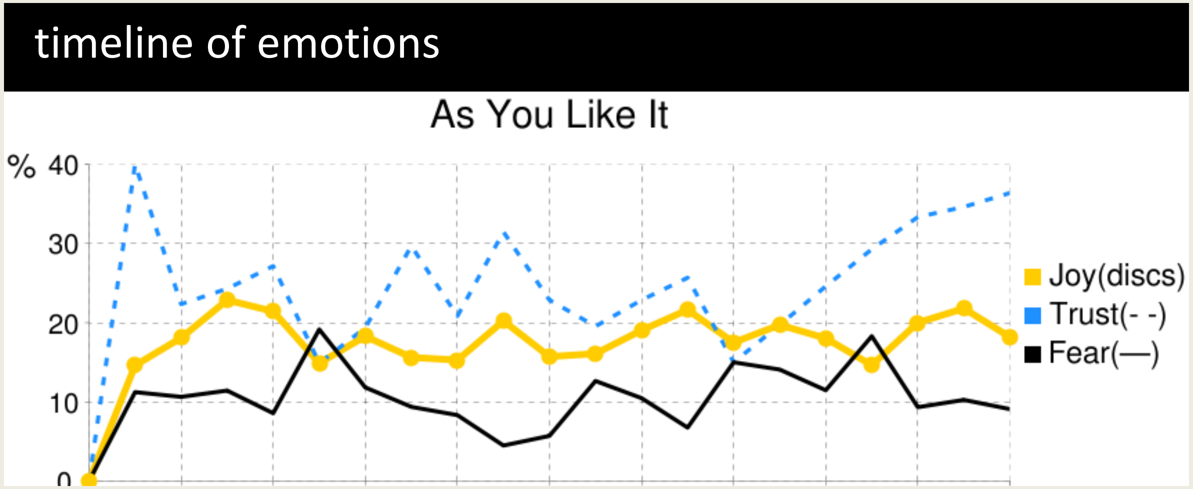
S. Mohammad (2011)

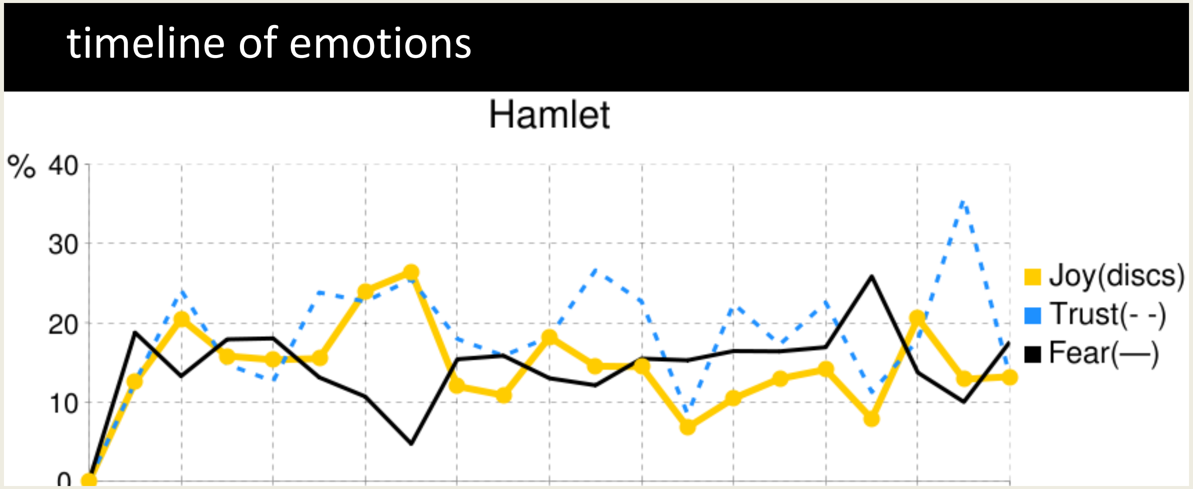
Difference in percentage scores for each of the eight basic emotions



Mining *Hamlet* and *As you like it (2)*

S. Mohammad (2011)





Things to solve when counting words

- Spelling errors (variation):
  - *comunism, Frakenstein*: do you mean xyz?
- Stemming:
  - *burning, burns, burnt and burned:* burn*
- Ambiguity/Synonymy (via natural language processing tools)
  - Ambiguity:
    - *bass (*fish, musical instrument, voice, …): ?
    - *fly (*Noun or Verb): ?
  - Synonymy:
    - *car, sedan, vehicle: ?*

More things to solve when counting words

- Distinction between function words (stopwords) and content words
- Relative frequency: document *versus* corpus
- Generic patterns in frequency: Zipf's Law

Variation in text mining

- Trends in sentiment: online reviews, news, etc.
- Cultural dynamics: changes in frequencies (http://www.culturomics.org/)
- Author features: gender, age, class, style, region, …
- Emergence of novel concepts/co-occurrences
- Topic-specific patterns: topic classification, topic clustering
- Patterns in online conversations
- Correlation studies

Algorithms, tools, resources

- WordNet (language-specific: Cornetto (NL)
- Wikipedia: source for disambiguation
- Ontologies
- LDA-framework (topic clustering)
- Training data for machine learning algorithms:
  manual annotation, Mechanical Turk
- Training corpora – annotation – MechTurk, etc/
- Validated toolboxes: xTas, LIWC, Stanford NLP, etc.
- getNgrams.exe to get n-gram data (also for Python)
- Coding languages: R, Python
- LinkedIn – group
- http://tedunderwood.com/2012/08/14/where-to-start-with-text-mining/ (widely cited blog)

Recent list of LinkedIn themes

- Data science shows surveys may assess language more than attitudes
- List of 50+ Machine Learning APIs - Mashape Blog
- List of 25+ Natural Language Processing APIs - Mashape Blog
- The Role of Text Mining in the Insurance Industry

What TM brings

- Structure for weakly structured data
- Data reduction: summarization
- Analysis focussed on specific aspects, e.g. named entities: person, location, organisation

  demo: Fido
- Tools for distant reading (versus close reading)

What helps TM to support you better

- Data, data, data (without serious size, analysis results are meaningless)
- Result visualization tools
- Understanding of how word counts relate to real life phenomena
- Research framework with quantitative perspective
- Programming skills?