Language models using ngrams and neural networks for text mining and computational social science

CSE 634 : Data Mining Concepts and Techniques
Prof. Anita Wasilewska

References

1. https://lifehacker.com/why-you-shouldnt-buy-a-new-book-on-amazon-1819404226
2. https://chandoo.org/wp/2013/07/01/introduction-to-excel-2013-data-model-relationships
3. http://slideplayer.com/slide/6183558/
4. http://feed140.com/helpful-twitter-bots/
5. https://www.google.com/search?q=amazon+reviews+images&source=lnms&tbm=isch&sa=X&ved=0ahUKEwj8ofjnsN7aAhVFiOKHaARCxAQ_AUICygC&biw=1280&bih=615#imgrc=xYm05HXg85L2HM:
6. The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining(1).-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011
7. https://homepages.inf.ed.ac.uk/lzhang10/slm.html
8. https://en.wikipedia.org/wiki/Language_model
9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419897/
10. https://en.wikipedia.org/wiki/Language_model
11. https://www.semanticscholar.org/paper/Text-mining-and-probabilistic-language-modeling-for-Lau-Liao/eaead47da83d850b2f5e0c60177db36e7cbc07b
12. CSE 634 Course Material - http://www3.cs.stonybrook.edu/~cse634/L1ch1introd.pdf
13. https://en.oxforddictionaries.com/explore/what-is-a-corpus
14. https://reutersinstitute.politics.ox.ac.uk/our-research/global-database-investigations-role-computer-assisted-reporter
15. https://en.wikipedia.org/wiki/N-gram#Applications_and_considerations
16. http://text-analytics101.rxnlp.com/2014/10/all-about-stop-words-for-text-mining.html
17. https://blog.bitext.com/what-is-the-difference-between-stemming-and-lemmatization/
18. https://www.gutenberg.org/ebooks/98
19. https://machinelearningmastery.com/gentle-introduction-bag-words-model/
20. http://www.holidayeducationist.com/how-to-build-kids-vocabulary/
21. http://textanalysisonline.com/nltk-wordnet-lemmatizer
22. http://www.aleron.org/our-services-and-expertise/transformation/

https://lifehacker.com/why-you-shouldnt-buy-a-new-book-on-amazon-1819404226

Overview

- We first set out to attempt to explain a deceptively simple yet powerful model that proved to be effective in text classification about ten years ago and paved the way for all other models. These are the n-gram language models.They are still used today even in the days of complex neural architectures.
- Then we will present one neural network based word embedding model which is now the state of the art and is used in many downstream text mining and NLP tasks.
- Finally we will put forward an application oriented research paper which uses very simple methods for language analysis but exhibits their power to mine insights of culture.

Why language models in Data mining

- Main aim is to extract knowledge from data
- Availability of huge text data
- Important Information hidden in text data
- Sometimes it is not possible to collect information from questionnaire
  e.g Patient's clinical history
    - Human experience
    - Interviews
    - Reviews
    - Information in books

Type de roche	CaO+MgO	Al2O3	TiO2	Fe2O3*	MnO	MgO	CaO	Na2O	K2O	P2O5	S	Zn	Pb
R. carbonatées	58.55	0.04	0.001	0.05	0.01	24.52	34.03	0.03	0.02	0.05	72	2	
R. carbonatées	55.92	0.06	0.001	0.08	0.02	23.54	32.38	0.02	0.02	0.01	89	6	
R. carbonatées	54.33	0.16	0.010	0.31	0.05	22.64	31.69	0.03	0.02	0.21	258	7	8
R. carbonatées	53.28	0.04	0.001	0.05	0.01	22.16	31.12	0.03	0.01	0.05	139	6	
R. carbonatées	52.29	0.08	0.001	0.15	0.02	16.45	35.84	0.03	0.06	10.41	754	5	4
R. carbonatées	51.88	0.08	0.002	0.14	0.03	21.42	30.46	0.03	0.04	0.23	341	4	
R. carbonatées	51.54	0.04	0.001	0.08	0.02	21.52	30.03	0.02	0.01	0.01	110	14	
R. carbonatées	51.53	0.14	0.005	0.26	0.03	21.54	29.99	0.03	0.05	0.08	728	6	

Clinical history

History of Present Illness

The patient is a 55-year old Caucasian man who presented to an outside hospital with a chief complaint of abdominal pain and was

Sale ID	Time	Customer	Product ID	Quantity
S00001	12/1/2012 9:00:00 AM	C0001	P025	1
S00002	12/1/2012 9:05:58 AM	C0025	P025	3
S00003	12/1/2012 9:11:33 AM	C0010	P001	2

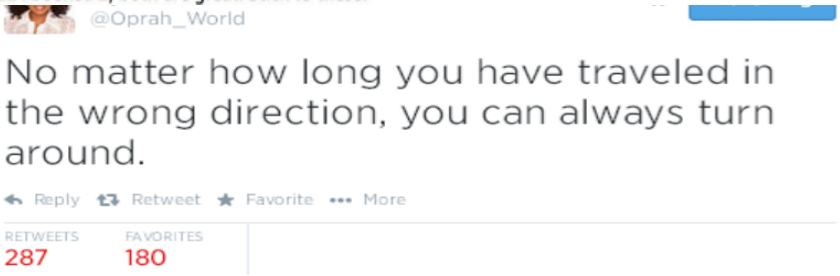
name	*age*	*income*	*loan_decision*
Sandy Jones	youth	low	risky
Bill Lee	youth	low	risky
Caroline Fox	middle_aged	high	safe
Rick Field	middle_aged	low	risky
Susan Lake	senior	low	safe
Claire Phips	senior	medium	safe
Joe Smith	middle_aged	high	safe
...	...	...	...

★★★★★ **These work!!**

By Covenant on August 29, 2017

Color: white9

When we lost the original adapter, we tried several ports: one headset and the other for another lightning port (for a fee). Novels - they do not work, even just headphones jacks. We return to these are their own work often have a good sound quality. They reach the original apple wrap (or at least a very good copy) and work like the original. I booked 2, both are great. Stick to these!



https://chandoo.org/wp/2013/07/01/introduction-to-excel-2013-data-model-relationships/
http://slideplayer.com/slide/6183558/
http://feed140.com/helpful-twitter-bots/
https://www.google.com/search?q=amazon+reviews+images&source=lnms&tbm=isch&sa=X&ved=0ahUKEwj8ofjnsN7aAhVFiOAKHaARCxAQ_AUICygC&biw=1280&bih=615#imgrc=xYm05H

What is Statistical Language Modeling (SLM)

A statistical language model (SLM) is a probability distribution P(s) over strings S that attempts to reflect how frequently a string S occurs as a sentence.

Common SLM techniques:

- N-gram model (most widely used SLM today.)
- Exponential language model
- Neural language model

https://homepages.inf.ed.ac.uk/lzhang10/slm.html
https://en.wikipedia.org/wiki/Language_model

Applications

- Information retrieval
- Google search sequence prediction
- detecting and analyzing fake reviews (i.e., spam)
- Identify Relevant New Information in Inpatient Clinical Notes

Many more:

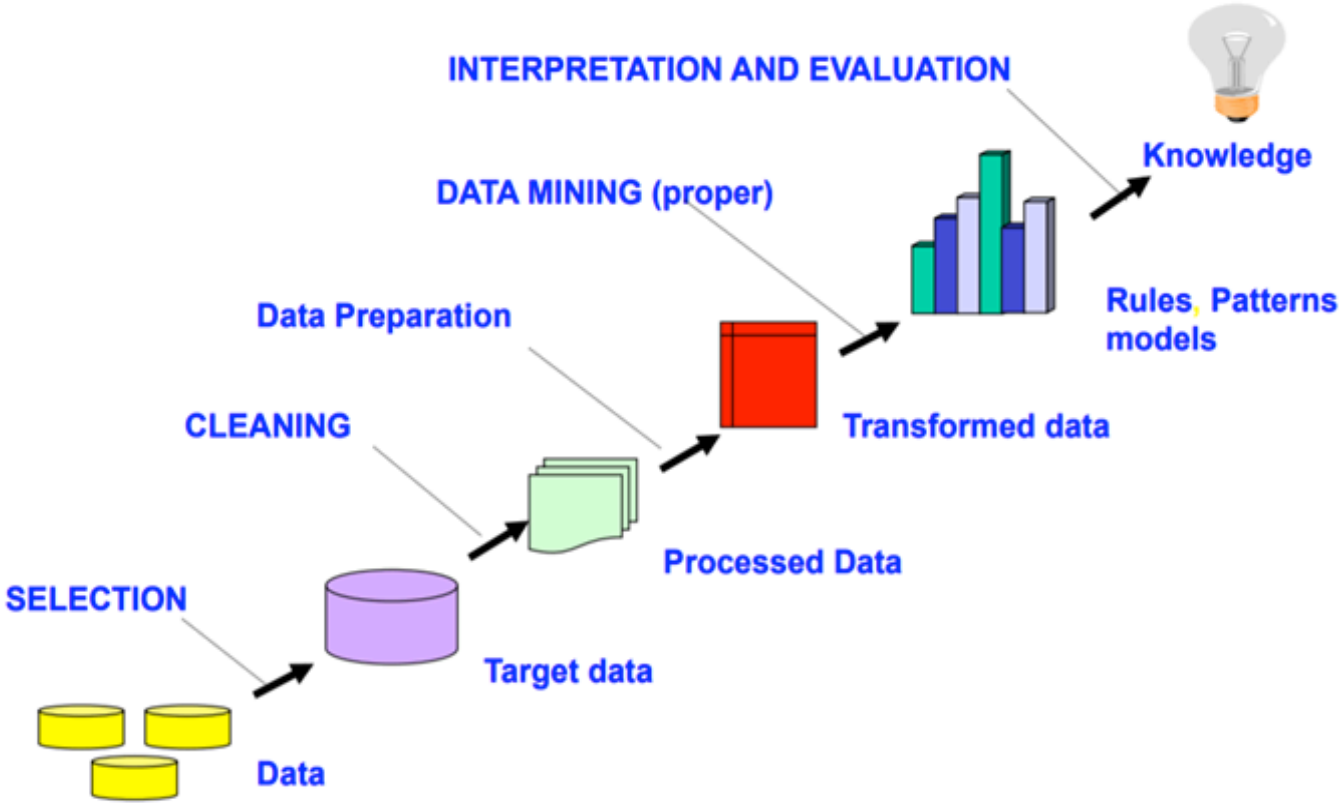
- speech recognition,
- machine translation,
- part-of-speech tagging,
- parsing,
- handwriting recognition

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419897/
https://en.wikipedia.org/wiki/Language_model
https://www.semanticscholar.org/paper/Text-mining-and-probabilistic-language-modeling-for-Lau-Liao/eaead47da83d850b2f58e0c60177db36e7cbc07b

Pre-Processing

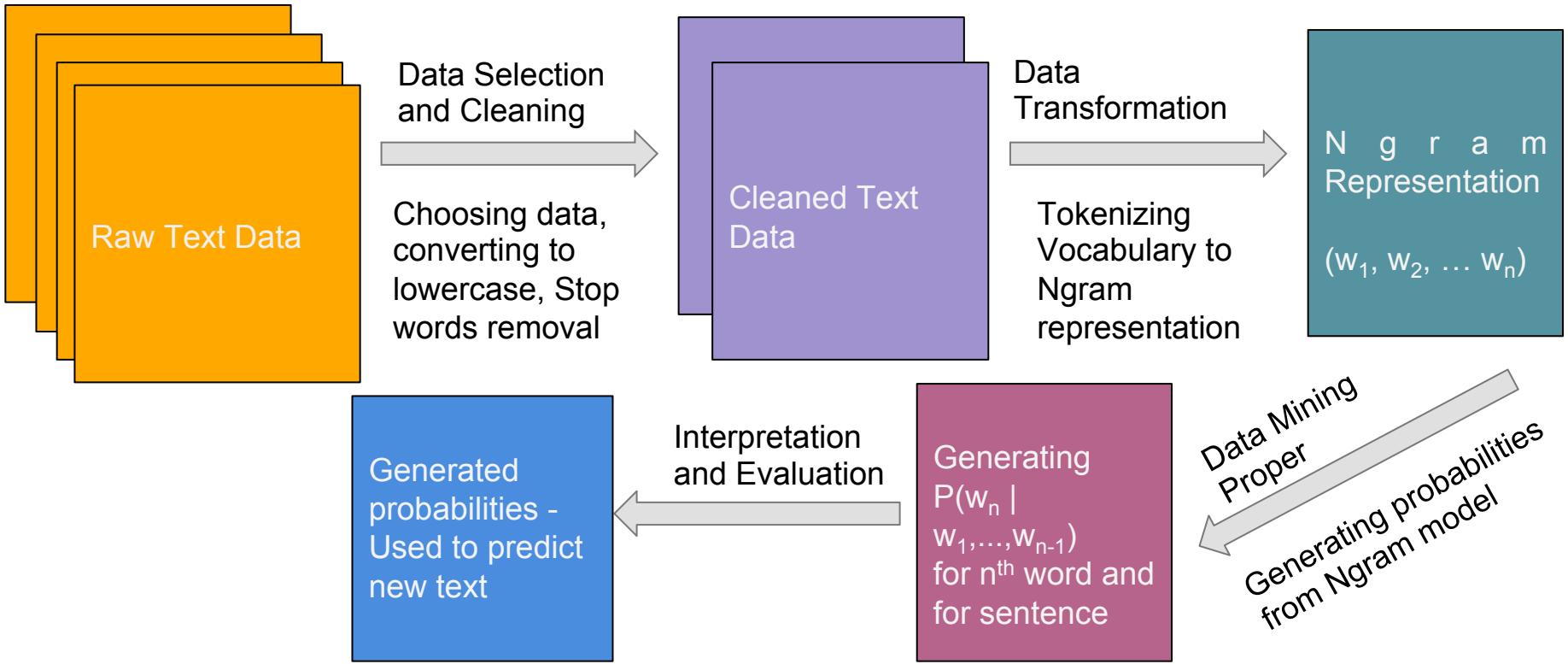
Preparing the data to be used for Ngrams

Data Mining Process - How it all fits in?



CSE 634 Course Material - http://www3.cs.stonybrook.edu/~cse634/L1ch1introd.pdf

Breaking it down



What is our Data?

- Usually in Text Mining, Corpus is used, which is a large collection of texts.
- It contains not only published works in which the text has been edited but also unpublished and unedited writing like emails and blogs.

https://en.oxforddictionaries.com/explore/what-is-a-corpus
https://reutersinstitute.politics.ox.ac.uk/our-research/global-database-investigations-role-computer-assisted-reporter

Let’s preprocess !!

- Ignoring case - Generally, all words converted to lowercase.
- Ignoring punctuation - done to trigger functionality.
- Collapse white spaces to single space while preserving paragraph marks, as they introduce layout or presentation and is not needed during prediction.
- Tokenize the sentence into words/phrases

https://en.wikipedia.org/wiki/N-gram#Applications_and_considerations

Stop the “Stop Words”

- Stop words are extremely frequent words which don’t contain much information and are removed. They can be general or domain specific.

- **General** - For example : Determiners (“the”, “an”, “a”) , Prepositions (in, under, towards, before), etc.

- **Domain Specific** - For example, in clinical texts, terms like “mcg” “dr.” and “patient” occur almost in every document that you come across. So, these terms may be regarded as potential stop words for clinical text mining and retrieval. Whereas terms such as ‘heart’, ‘failure’ and ‘diabetes’ are more important.

http://text-analytics101.rxnlp.com/2014/10/all-about-stop-words-for-text-mining.html

Stemming and Lemmatization

- Reducing the inflectional forms of words into a common base or root.
- **Stemmer** cuts the end or beginning of word taking into account common prefixes or suffixes.
- **Lemmatizer** considers morphological analysis of the words. “Lemma is the base form of all its inflectional forms, whereas a Stem isn’t.”

Form	Suffix	Stem
studies	-es	studi
studying	-ing	study
niñas	-as	niñ
niñez	-ez	niñ

Form	Morphological information	Lemma
studies	Third person, singular number, present tense of the verb study	study
studying	Gerund of the verb study	study
niñas	Feminine gender, plural number of the noun niño	niño
niñez	Singular number of the noun niñez	niñez

https://blog.bitext.com/what-is-the-difference-between-stemming-and-lemmatization/

Consider this

Consider the following snippet of the first few lines of text from the book “A Tale of Two Cities” by Charles Dickens, taken from Project Gutenberg :

It was the best of times,

it was the worst of times,

it was the age of wisdom,

it was the age of foolishness,

Let’s treat each line as a document and the 4 lines are the corpus of documents.

https://www.gutenberg.org/ebooks/98
https://machinelearningmastery.com/gentle-introduction-bag-words-model/

Vocabulary is ...

We make a list of all of the words in our model vocabulary by tokenizing. The unique words here are (ignoring case and punctuation):

- “it”
- “was”
- “the”
- “best”
- “of”
- “times”
- “worst”
- “age”
- “wisdom”
- “foolishness”

This is a vocabulary of 10 words from a corpus containing 24 words.

http://www.holidayeducationist.com/how-to-build-kids-vocabulary/

Vocabulary is ...

After Lemmatizing and stop words removal, we get

- “it”
- “best”
- “time”
- “bad”
- “age”
- “wisdom”
- “foolishness”

http://textanalysisonline.com/nltk-wordnet-lemmatizer

Transformation

- Cleaning reduces the vocabulary of words. The vocabulary generated from cleaning is converted to Ngram representation by grouping n words into a sequence.
- Each word is called as a token or gram.

http://www.aleron.org/our-services-and-expertise/transformation/

N-grams

For example, a vocabulary of two-word pairs is called “bigram model”. Three word sequence models are called trigrams. The bigrams in the first line of our example are:

- “it best”
- “best time”

The general approach of creating sequence of n-words is called n-gram model.

Language Modeling

- A language model is something that specifies the following two quantities, for all words in the vocabulary (of a language).

  1. Probability of a sentence or sequence

     $$Pr(w1, w2, \ldots, wn) = \#(w1, w2, \ldots, wn) / N$$

     - e.g., Pr(I, love, fish) = Pr(W1 = I, W2 = love, W3 = fish)
     - Disadvantages: We add and invent words … web is also no match for our creativity

  2. Probability of the next word in a sequence

     $$Pr(wk+1 \mid w1, \ldots, wk) = \#\,(\,wk+1, wk\,,\ldots,\,w1) / \#\,(\,wk, wk\text{-}1, \ldots\,,\,w1)$$

     - Disadvantages: Estimating conditional probabilities with long contexts is difficult...conditioning on 4 or more words itself is hard

https://web.stanford.edu/~jurafsky/NLPCourse.html

N-gram models

Markov Assumption

- Next event in a sequence depends only on its immediate past (context).

- N-grams
  - Unigram model Pr(wk+1) - Protein Sequencing (… A, G, C, T, T, G, A …)
  - Bigram model Pr(wk+1 | wk) - (… AG, GC, CT, TT, TG, GA …)
  - Trigram model Pr(wk+1 | wk-1, wk) - (… AGC, GCT, CTT, TTG, TGA, …)
  - 4-gram model Pr(wk+1 | wk-2, wk-1, wk) - (… AGCT, GCTT, CTTG, …)

Other contexts are possible and in many cases models tend to be more complex

https://web.stanford.edu/~jurafsky/NLPCourse.html

Efficient Estimation vs Generalization

• We can estimate unigrams quite reliably but they are often not a good model.

• Higher order n-gram require large amounts of data but are better models.

- However, they have a tendency to overfit the data.
- Issues with gram models
  - New words and sequence gets zero probability

Ex. Pr(a, an, the, a, an, 634) = Pr(CSE, 634, is, exciting) = 0

  - A related issue is too much mass for rare events

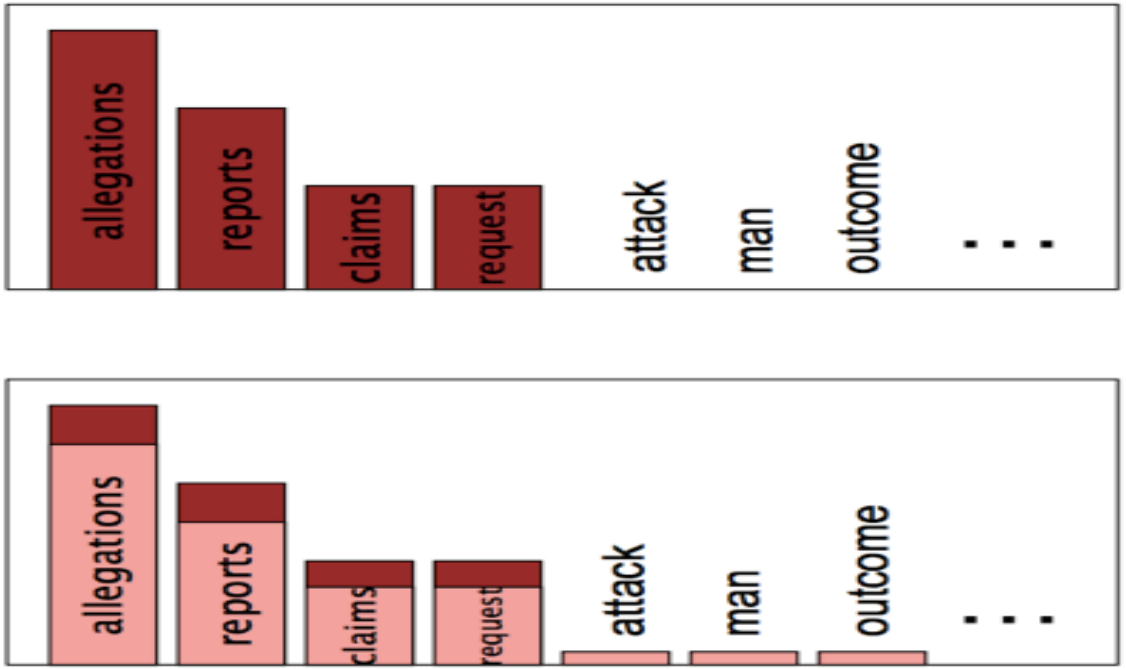
Ex. If the only sentence about CSE634 you have in the training data is:“CSE634 is awesome” then

Pr(awesome|CSE634, is) = ?

https://web.stanford.edu/~jurafsky/NLPCourse.html

Smoothing

- Pr(w|Denied)



https://web.stanford.edu/~jurafsky/NLPCourse.html

Laplace Smoothing

- Assume that there were some additional documents in the corpus, where every possible sequence of words was seen exactly once
- For bigrams, this means that every possible bi-gram was seen at least once.
- As k increases, it l minative natures erodes

$$Pr_L(w_i|w_{i-1}) = \frac{count(w_{i-1}, w_i) + 1}{count(w_{i-1}) + V}$$

$$Pr_{L_k}(w_i|w_{i-1}) = \frac{count(w_{i-1}, w_i) + k}{count(w_{i-1}) + kV}$$

https://web.stanford.edu/~jurafsky/NLPCourse.html

Good Turing - Smoothing Technique

- Notation:

  Nc = Frequency of frequency c (Count of things we have seen c times)

- Ex. Data mining is not data mining but is mining of knowledge

3	2	2	1	1	1	1
Mining	Data	Is	Knowledge	But	Not	Of

N1	N2	N3
4	2	1

https://web.stanford.edu/~jurafsky/NLPCourse.html

Intuition - Good Turing

- Assuming you are fishing and caught:

  10 carp, 3 perch, 2 whitefish, 1 trout, 1 salmon, 1 eel = 18 fish

- How likely is it that next species is trout?

  1/18

- How likely is it that next species is new (i.e. catfish or bass)
- Let's use our estimate of things-we-saw-once to estimate the new things.

  3/18 (because N1=3)

- Assuming so, how likely is it that next species is trout?

  Must be less than 1/18

https://web.stanford.edu/~jurafsky/NLPCourse.html

Calculation - Good Turing

P^*_{GT}(things with c frequency) = c^* / N

$$P^*_{GT}(\text{things with zero frequency}) = \frac{N_1}{N}$$

- Unseen (bass or catfish)
  - c = 0:
  - MLE p = 0/18 = 0
  - P^*_{GT} (unseen) = N_1/N = 3/18

$$c^* = \frac{(c+1)N_{c+1}}{N_c}$$

- Seen once (trout)
  - c = 1
  - MLE p = 1/18
  - C*(trout) = 2 * N2/N1
    = 2 * 1/3
    = 2/3
  - P^*_{GT}(trout) = 2/3 / 18 = 1/27

https://web.stanford.edu/~jurafsky/NLPCourse.html

Evaluation of n-gram and Perplexity

- How well can we predict the next word?
- It is the inverse probability of the test set, normalized by the number of words
- For probability of sentence:

$$PP(W) = P(w_1 w_2 ... w_N)^{-\frac{1}{N}}$$

$$= \sqrt[N]{\frac{1}{P(w_1 w_2 ... w_N)}}$$

- For Bigram:

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_{i-1})}}$$

https://web.stanford.edu/~jurafsky/NLPCourse.html

Example of Perplexity

- Lower Perplexity = Better the model
- Training 38 million words, test 1.5 million words, WSJ

N- gram	Unigram	Bigram	Trigram
Perplexity	962	170	109

Pros and cons n-grams

- Pros
  - Very simple to understand
  - Training is very fast compared to NN
  - Efficient use of statistics
- Cons
  - Not Scalable
  - It doesn’t capture inherent understanding of language like synonyms, antonyms etc.
  - Disproportionate importance to large counts
  - Complex interactions between words are not captured . Ex. Relation between state and capital

Neural Language Models using Glove Word Embeddings

- Developed by Chris Manning and Richard Socher at Stanford(2014),the main intuition underlying the model is the simple observation that ratios of **word-word co-occurrence probabilities have the potential for encoding some form of meaning.**
- Thus the authors of the model design a neural network(NN) which minimizes the difference between the input one hot vector representations of a pair of words and their value in the co-occurrence matrix.
- The NN also reduces the dimensionality of the very large one-hot form to a more dense representation by hidden layer transformation.

Co-occurrence matrix

Corpus = {"I like deep learning"
"I like NLP"
"I enjoy flying"}

Context = previous word and next word

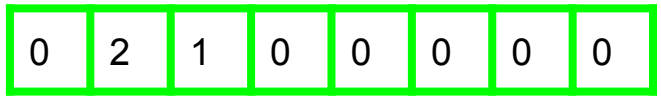
counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

nlp.cs.tamu.edu/resources/wordvectors.ppt

Can’t the rows/columns be vectors?

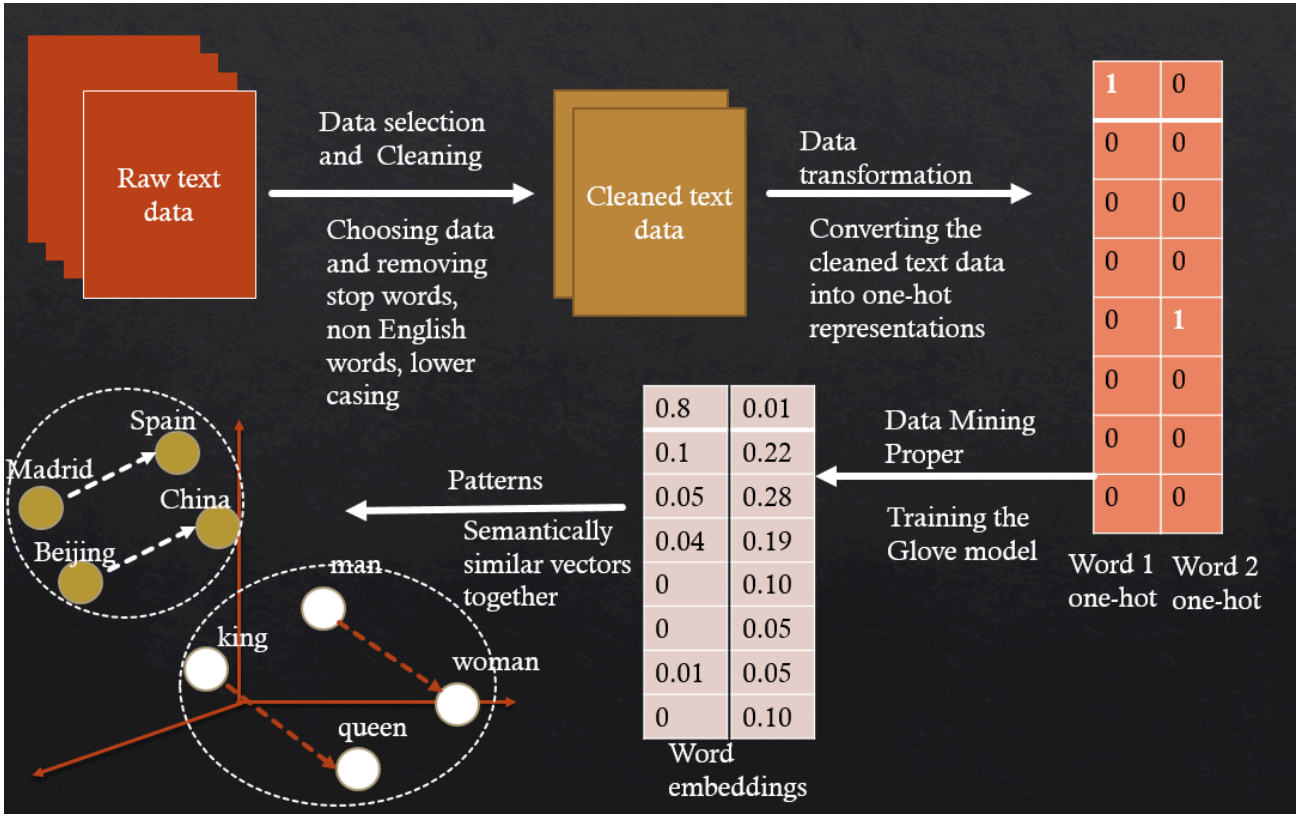
- Ideally each of the rows or columns could be word vectors, but the extensive vocabulary will provide a very sparse matrix dominated by 0 elements.
- Also the size of each vector will be exceptionally huge given the vocabulary size in English.
- Moreover we produce words everyday, like “Selfie” or “carpooling”

counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0



Vector for I from this co-occurrence matrix

Overview


Raw text data
Data selection and Cleaning
Choosing data and removing stop words, non English words, lower casing
Cleaned text data
Data transformation
Converting the cleaned text data into one-hot representations
1 0
0 0
0 0
0 0
0 1
0 0
0 0
0 0
Word 1 one-hot
Word 2 one-hot
Data Mining Proper
Training the Glove model
0.8 0.01
0.1 0.22
0.05 0.28
0.04 0.19
0 0.10
0 0.05
0.01 0.05
0 0.10
Word embeddings
Patterns
Semantically similar vectors together
Spain
Madrid
China
Beijing
man
king
woman
queen

Objective function- key difference compared to word2vec:

$$J(\theta) = \frac{1}{2} \sum_{i,j=1}^{W} f(P_{ij})(u_i^T v_j - \log P_{ij})^2$$

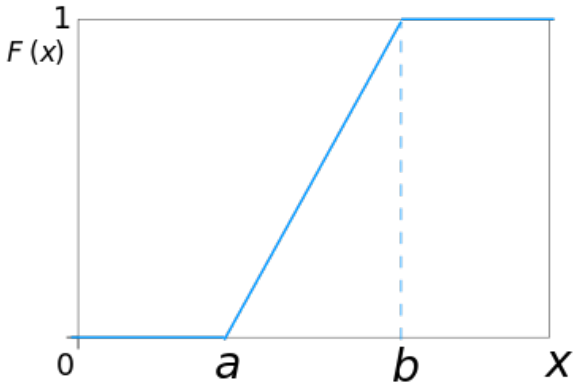
- θ here stands for all the parameters of the model which is the set of vectors u and v.
- P_{ij} stands for the co-occurrence matrix element corresponding to word in row i and word in column j.
- The log function plays a very important role. Since words like “I am” will often co occur with each other, there needs to be a cap on their frequencies which is what the log function does. (Log Smoothing)

nlp.cs.tamu.edu/resources/wordvectors.ppt

Objective function

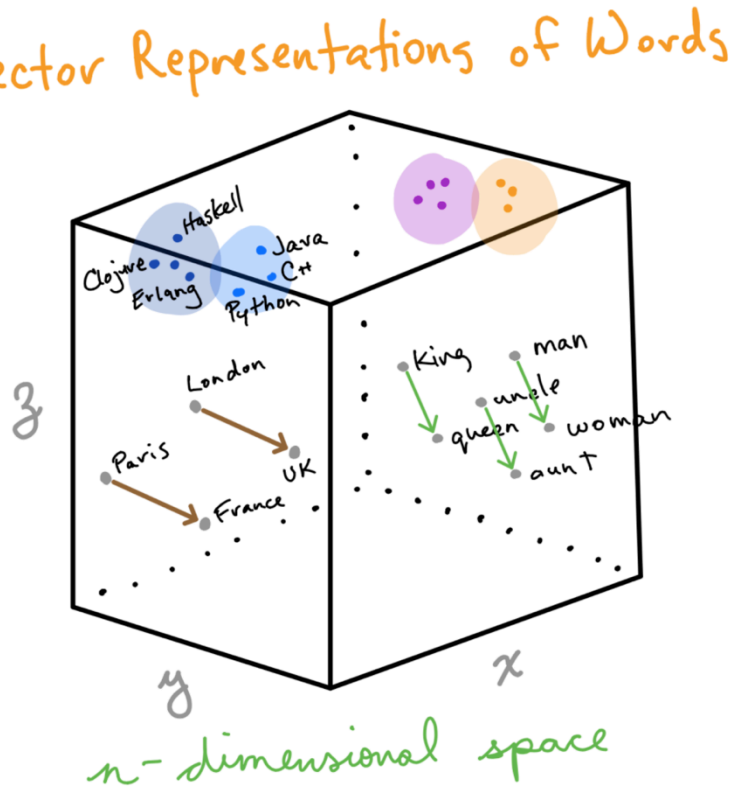
$$J(\theta) = \frac{1}{2} \sum_{i,j=1}^{W} f(P_{ij})(u_i^T v_j - \log P_{ij})^2$$

- Another important aspect of the objective function is the function $f(P_{ij})$
- Any value beyond the threshold will get “clipped” to that threshold thereby weighing in the importance based on occurrence but not over stating it or biasing it.

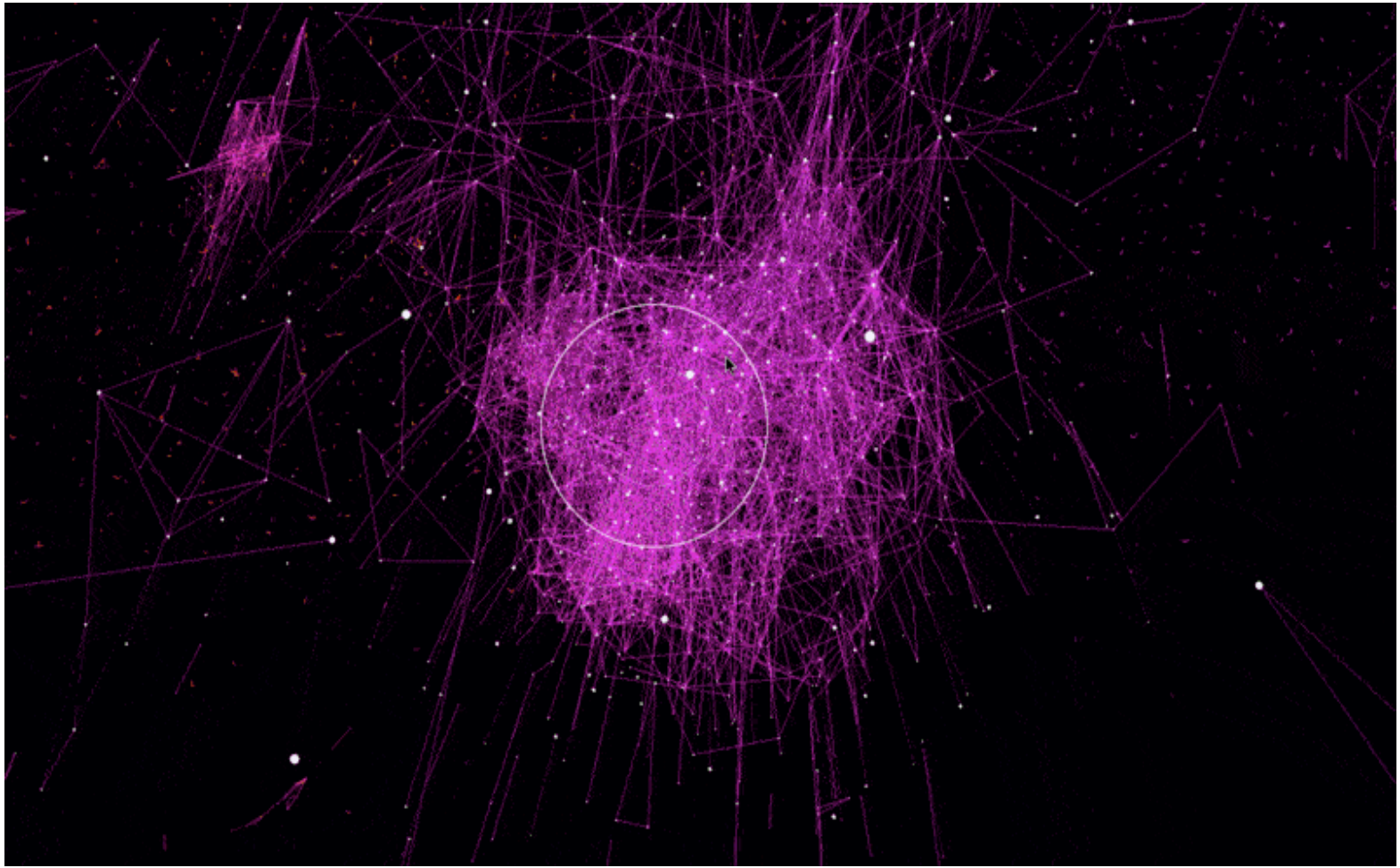


nlp.cs.tamu.edu/resources/wordvectors.ppt

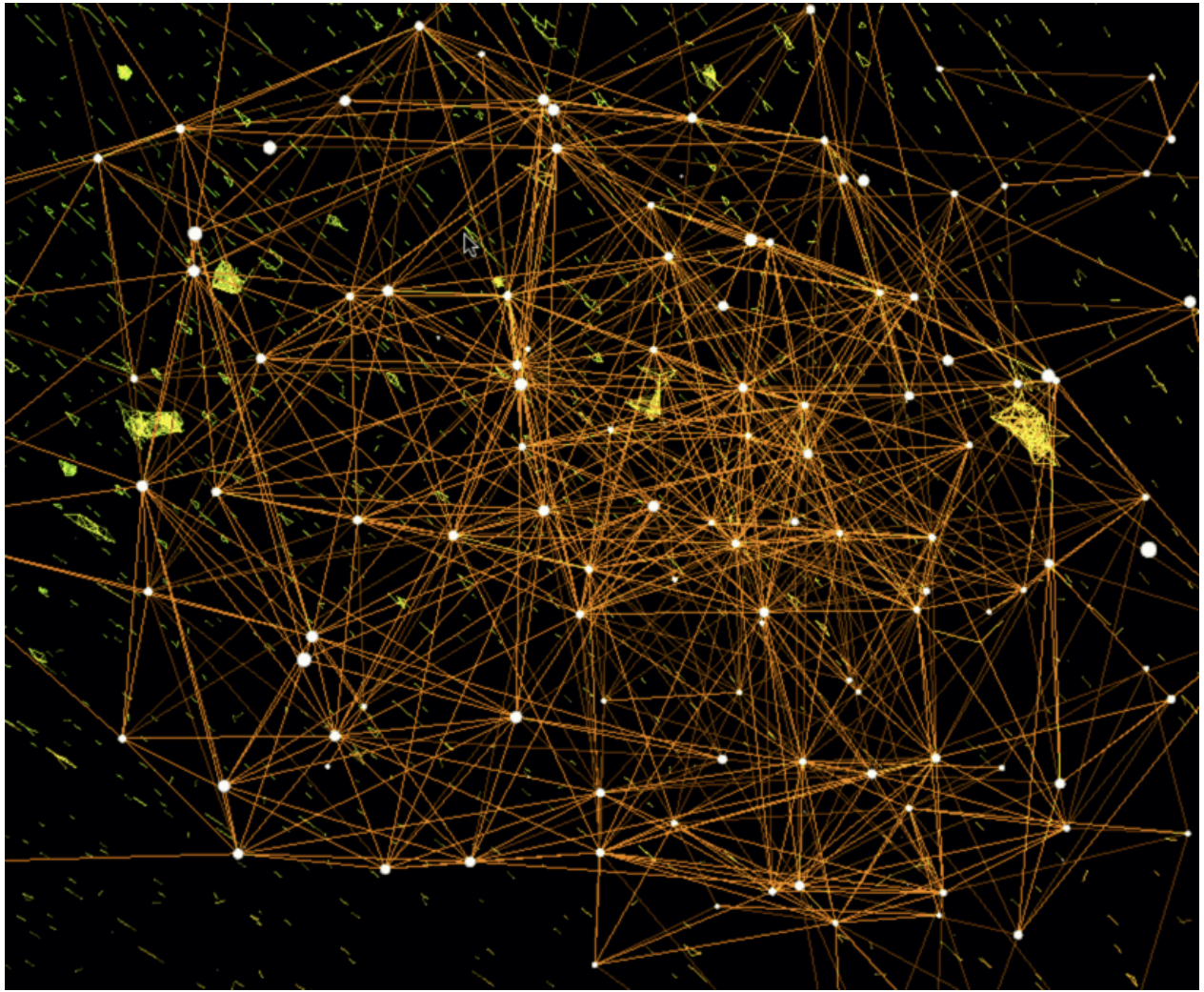
Glove vectors mapped in word space



https://www.quora.com/What-does-the-word-embedding-mean-in-the-context-of-Machine-Learning



reference



reference

Advantages of Glove model

- Can adapt to any kind of sentences and overcome sparsity
- Resulting representations have semantic similarity
- Being probabilistic is nature, it is supposed to perform superior to deterministic methods
- More intuitive and direct as compared to word2vec which does a proxy task

Disadvantages of Glove model

- Inability to handle unknown or Out Of Vocabulary words
- Scaling to new languages requires new embedding matrices
- Extremely long training time

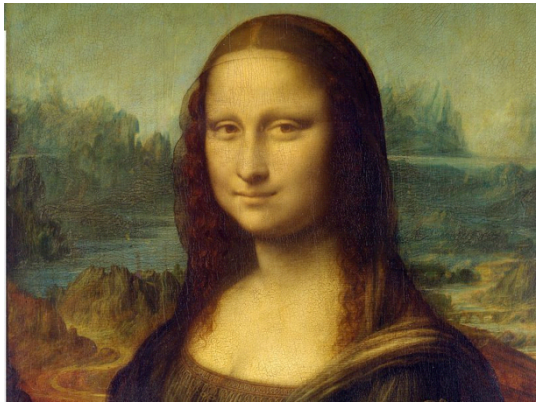
Michel Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg et al.

"Quantitative analysis of culture using millions of digitized books."

Science Journal 14 Jan 2011:
Vol. 331, Issue 6014, pp. 176-182
DOI: 10.1126/science.1199644

The Goal: Culture Mining

- To understand human culture of the past and now and see how it has changed over the course of time.
- All this while quantitative methods into the study of culture had been hampered by the lack of suitable data.
- The authors thus put forward Cultormics-. Computational analysis enabled methods to observe cultural trends and subject them to quantitative investigation



https://www.smithsonianmag.com/smart-news/

https://www.pinterest.co.uk/pin/492792384209844159/

The idealistic method

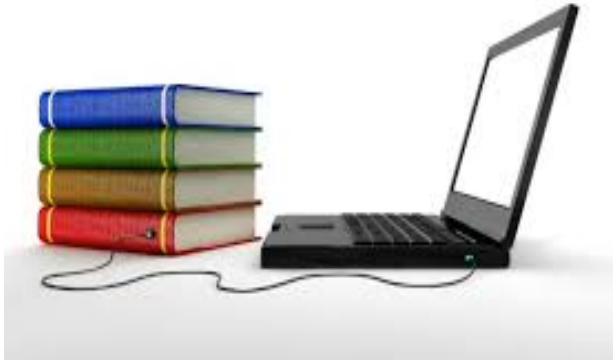
They authors decide to read all the books that have been written from the 1800 to 2000.

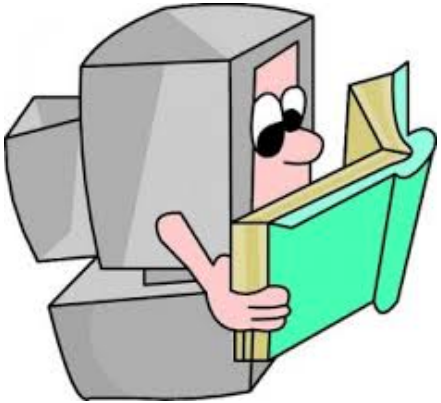
But there's an issue.

https://www.dreamstime.com/editorial-stock-image-loads-books-barcelona-spain-january-bookshelves-street-market-january-barcelona-spain-outdoor-book-markets-common-image46236724

https://www.google.com/imgres?imgurl

The practical method



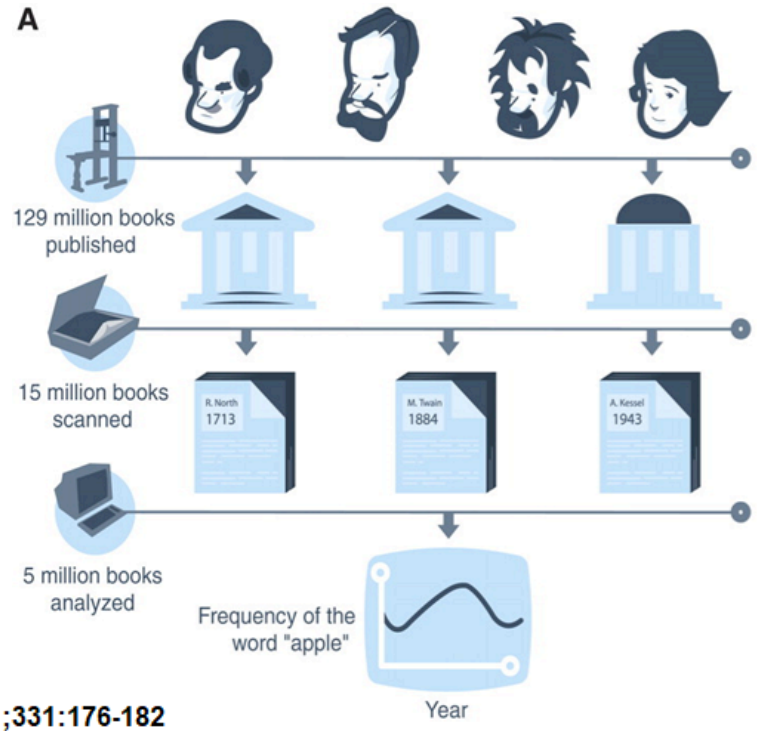


http://www.full-stop.net/2011/11/10/blog/kerem-ozkan/get-over-it-why-you-need-to-start-reading-on-your-computer/

http://www.teachingcollegeenglish.com/2010/09/page/4/

Jean-Baptiste Michel et al. Science 2011;331:176-182

The practical method



Jean-Baptiste Michel et al. Science 2011;331:176-182

The data

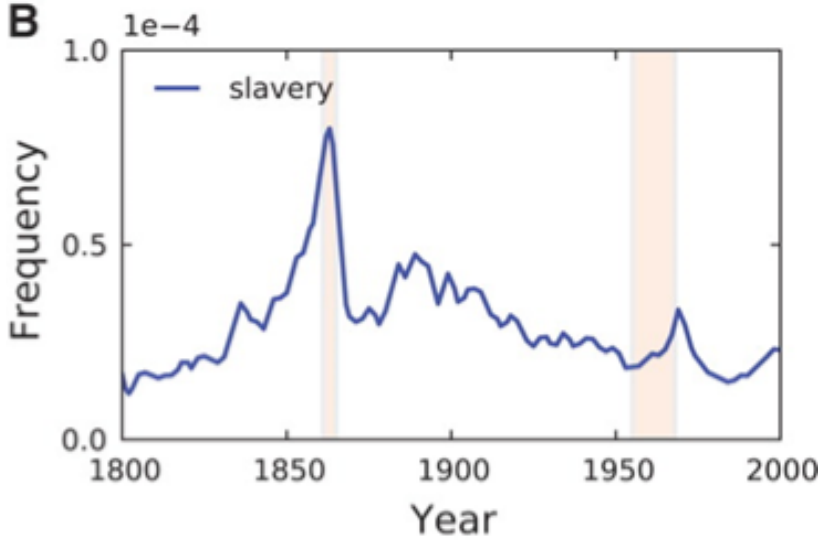
- Most books were drawn from over 40 university libraries around the world. Each page was scanned and the text was digitized by means of optical character recognition (OCR)
- Metadata describing the date and place of publication were provided by the libraries and publishers
- Authors selected 5,195,769 digitized books containing ~4% of all books ever published based on the overall quality of the digitized copies
- The corpus contains over 500 billion words, in English (361 billion), French (45 billion), Spanish (45 billion), German (37 billion), Chinese (13 billion), Russian (35 billion), and Hebrew (2 billion)

What do they analyze

- To make release of the data possible in light of copyright constraints, they restricted this initial study to the question of how often a given 1-gram or n-gram was used over time.
- Main idea was usage frequency.
- **Usage frequency is computed by dividing the number of instances of the n-gram in a given year by the total number of words in the corpus in that year**
- Total table of 2 billion ngrams

How important is language to culture?

In 1861, the 1-gram “slavery” appeared in the corpus 21,460 times, on 11,687 pages of 1208 books. The corpus contains 386,434,758 words from 1861; thus, the frequency is 5.5 × 10−5 . The use of “slavery” peaked during the Civil War (early 1860s) and then again during the civil rights movement (1955–1968)



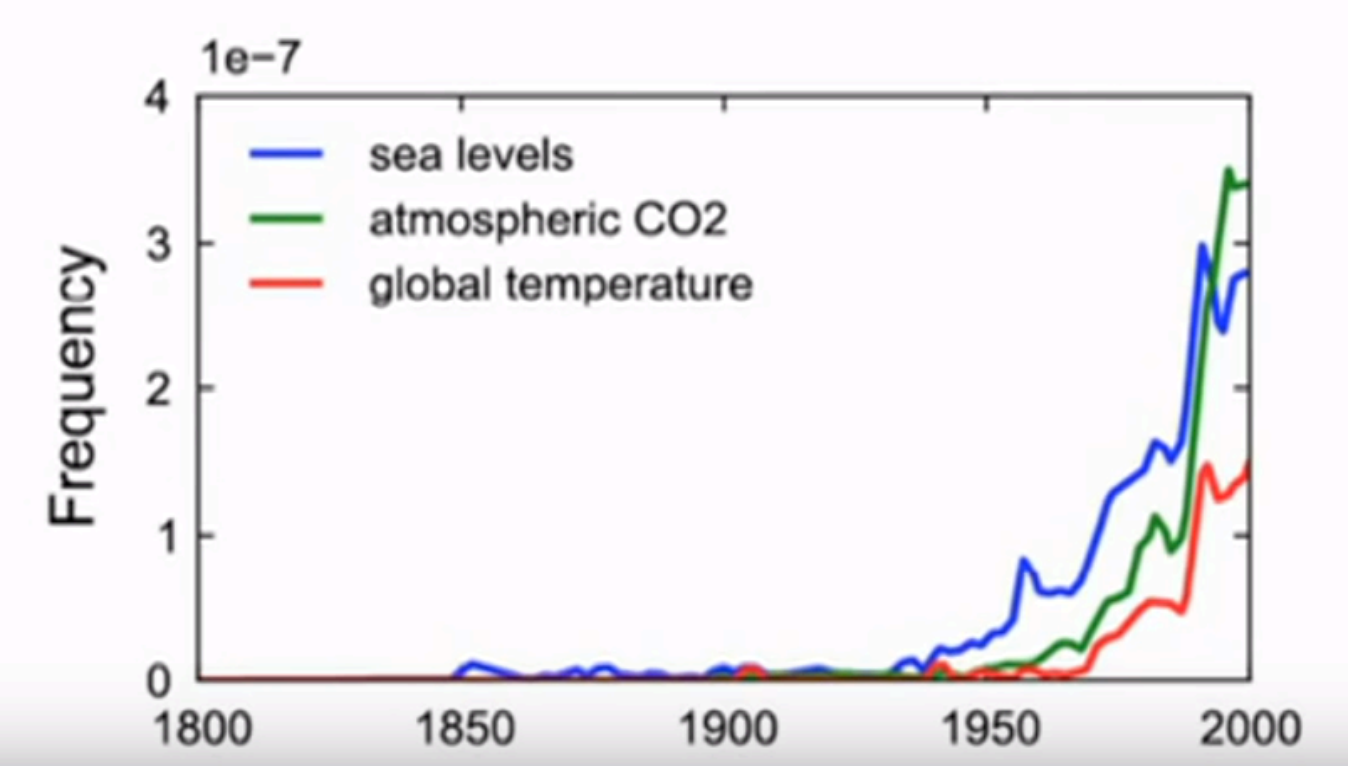
Jean-Baptiste Michel et al. Science 2011;331:176-182

How important is language to culture?

Frequency of terms “Great War” vs “World War I” and “World War II”. References to “the Great War” peak between 1915 and 1941. But although its frequency drops thereafter, interest in the underlying events had not disappeared; instead, they are referred to as “World War I”

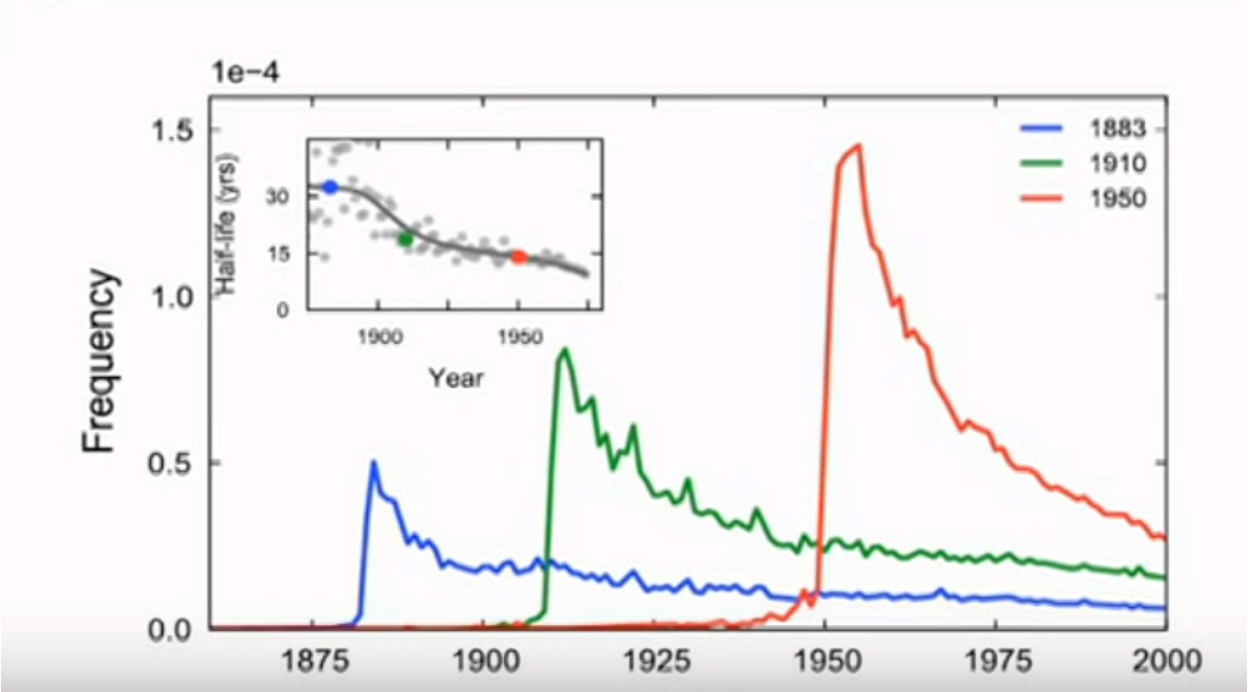
Jean-Baptiste Michel et al. Science 2011;331:176-182

Environment issues being raised



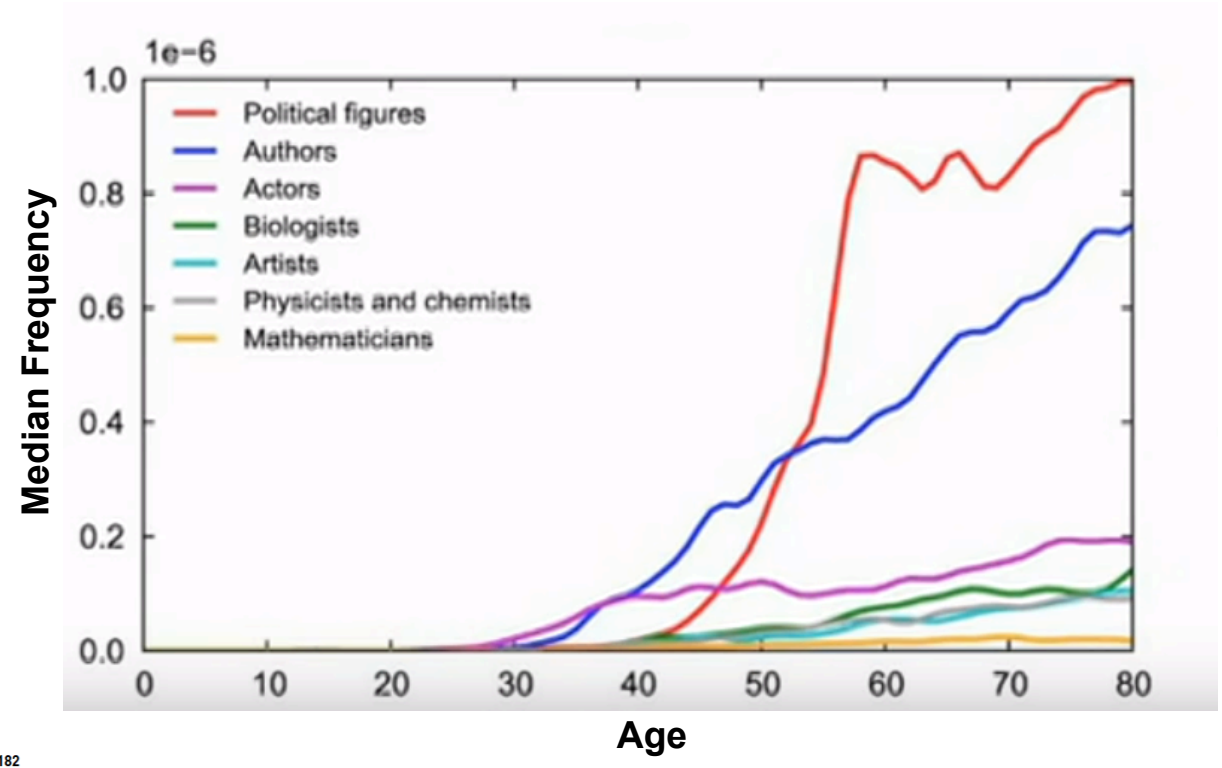
Jean-Baptiste Michel et al. Science 2011;331:176-182

Caring less about the past or do we really live in the present?



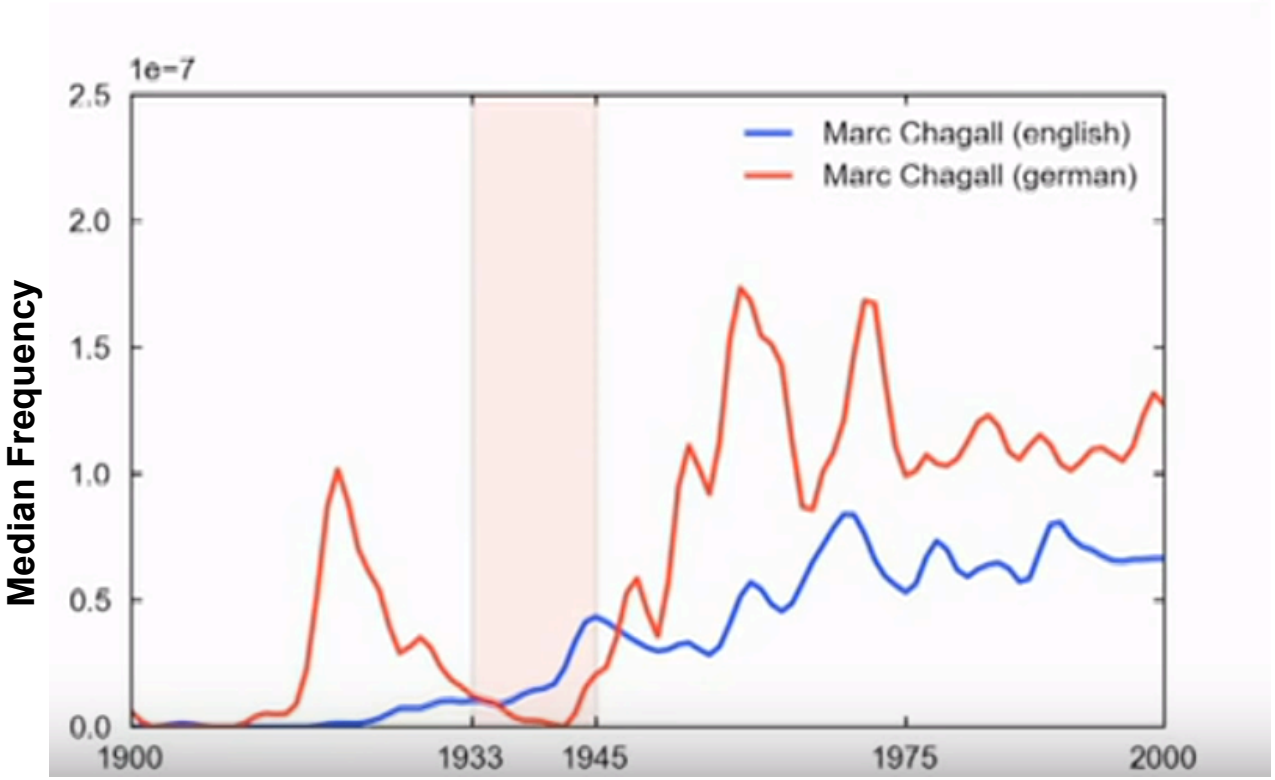
Jean-Baptiste Michel et al. Science 2011;331:176-182

Professions and fame



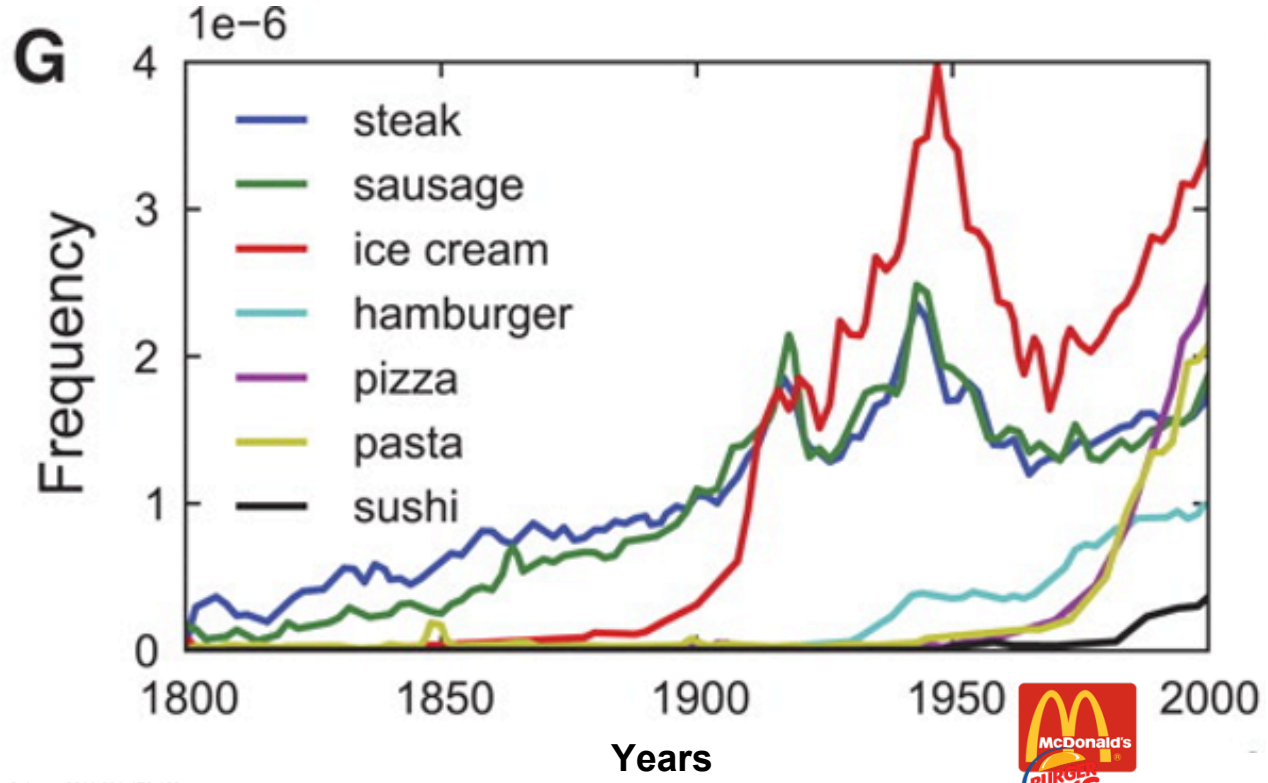
Jean-Baptiste Michel et al. Science 2011;331:176-182

Suppression trends



Jean-Baptiste Michel et al. Science 2011;331:176-182

Food Trends

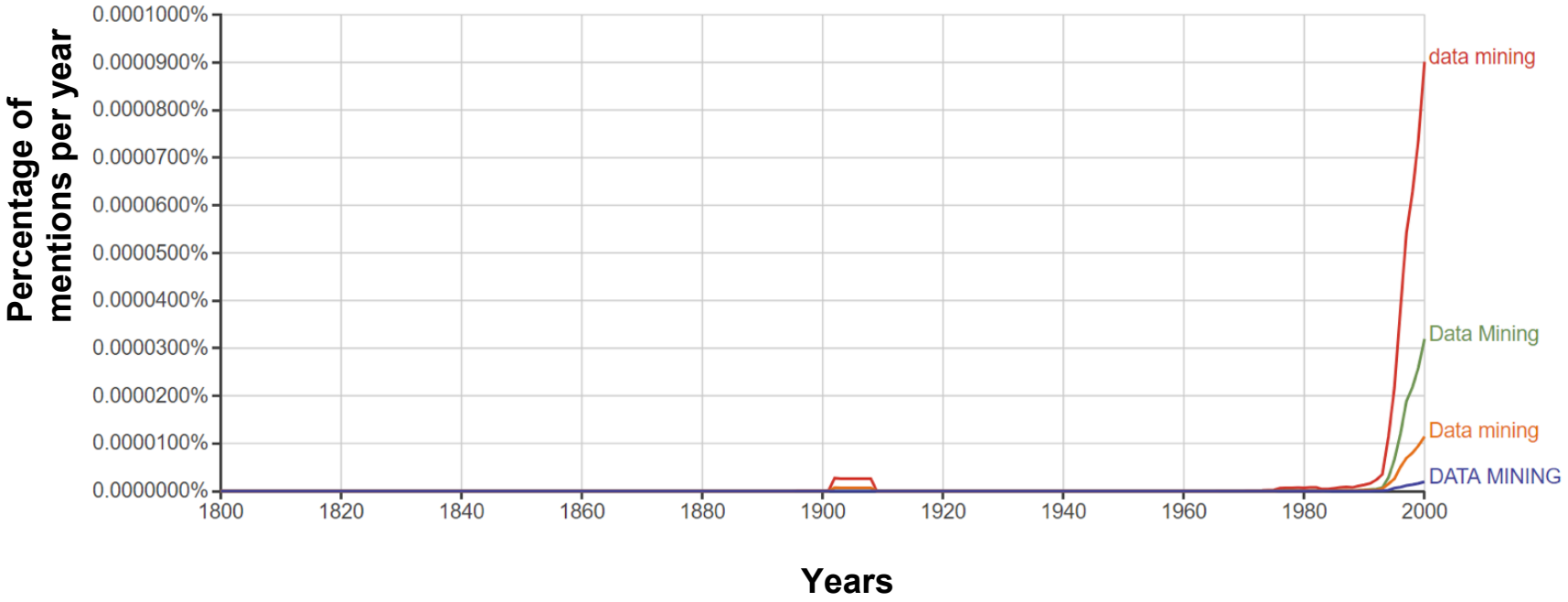


Jean-Baptiste Michel et al. Science 2011;331:176-182

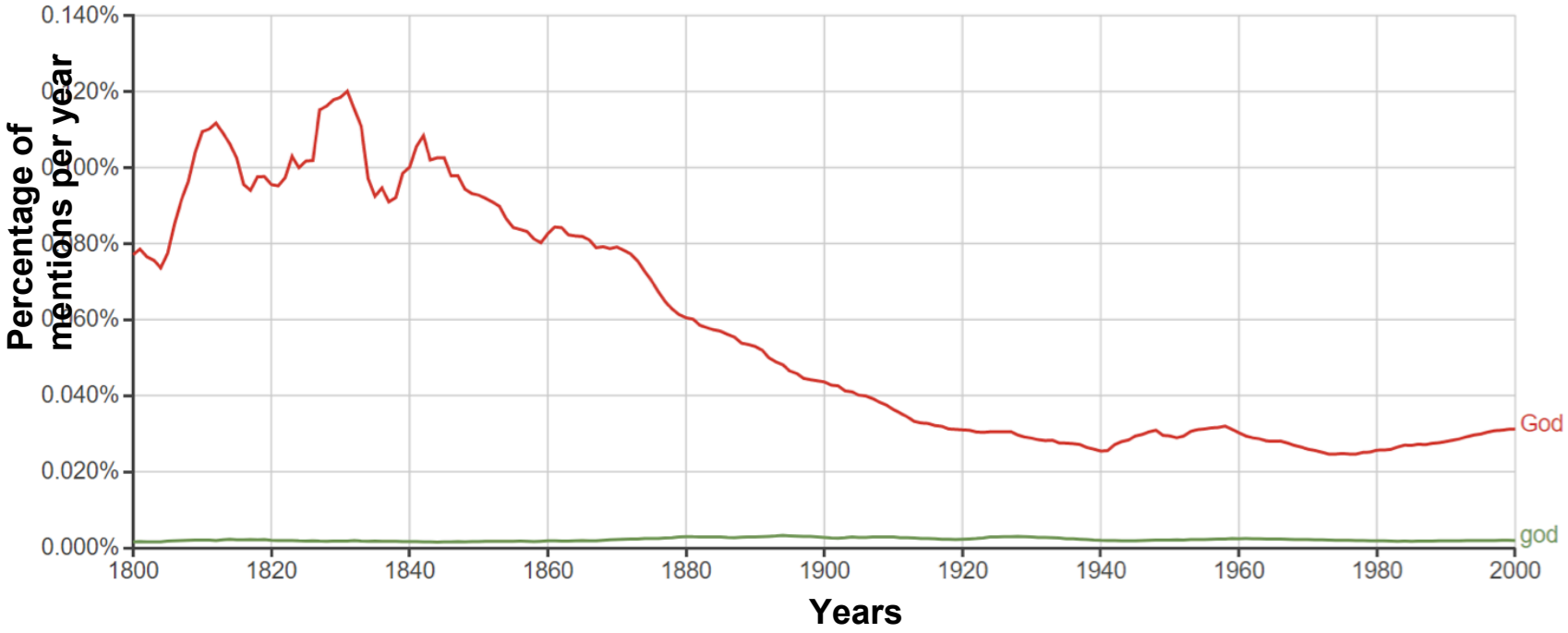
http://diylogodesigns.com/blog/mcdonalds-logo/

https://commons.wikimedia.org/wiki/File:Burger_King_Logo.svg

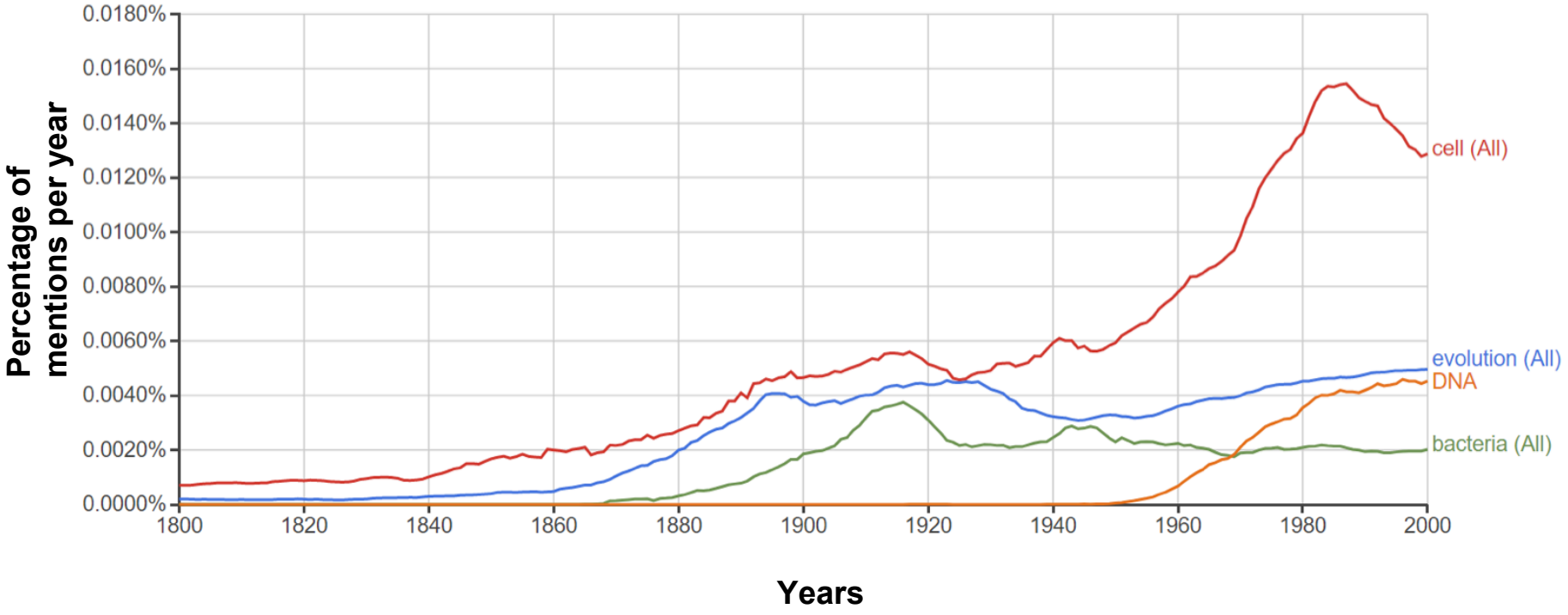
So when did data mining appear?



Theological plot



Evolution of biology



Thank You