Natural Language Processing in Python

Fiona Shen-Bayh

University of California, Berkeley

February 2017

Overview

- Natural language processing
  - Text as Data
  - Cleaning text
  - Converting text to numbers
  - Interpretation
- Common methods and applications

Text as Data

We know how to read and analyze a single document.

But what about hundreds or even thousands of documents?

Is it possible to automate this task?

Natural Language Processing

NLP is a process of using computer algorithms to understand and analyze natural language.

Everyday examples:

Autocorrect on your phone

Google translate

Spam filters in your email app

NLP and Social Research

As more text becomes available through the web and other digital databases, researchers are increasingly using NLP methods to analyze large bodies of text for social science and humanities applications.

NLP and Social Science

Grimmer (2010): devised topic model to measure political priorities of U.S. senators based on content in their press releases

Stewart and Zhukov (2009): developed a codebook of discourse for Russian political and military elites to measure whether actors have a restrained, activist, or neutral positions on the use of force

King, Pan, Roberts (2013): web scraped thousands of Chinese social media posts and analyzed content using NLP to identify patterns of state censorship

NLP Software

Many off-the-shelf programs have not been tested on a wide range of content types.

Writing your own NLP program gives you enhanced flexibility and control over your text analysis.

What NLP can do

Automatically extract important features from large bodies of text.

Provide efficient means of synthesizing and comparing different texts.

Classify texts into important groupings.

What NLP cannot do

A computer is not a human.

As of now, humans are still better interpreters of natural language than computers.

(But we are slower readers!)

Acquiring Text

Archives, libraries

Documents will need to be scanned and converted to TXT using OCR

Online databases

LexisNexis, ProQuest

Anywhere from the web

Download with web scraper or web driver

Terminology

Natural languages are used for everyday communication (English, Norwegian).

Artificial languages are used by programmers (Python, C++, JavaScript).

More Terminology

Unit of analysis: text or document

Body of texts: corpus

Bag-of-words: texts can be described by word frequencies, assumes that syntax does not matter.

Pre-Processing Text

Not all words are considered equally important to our analysis.

Pre-processing involves steps to “clean” the text of irrelevant data.

Pre-Processing Text

Pre-processing involves several steps to “clean” the text of irrelevant data.

punctuation: remove ?,.!-”()[]{}

Pre-Processing Text

Pre-processing involves several steps to “clean” the text of irrelevant data.

stemming: remove ends of words

Family, families, families’ → *famil*

Pre-Processing Text

Pre-processing involves several steps to “clean” the text of irrelevant data.

“stop” words: drop function words

it, the, a, is, are…

Creating a Vocabulary

After pre-processing, we are left with a *vocabulary* of unique words that will become the primary features of our text.

Transforming the Vocabulary

This vocabulary is transformed into vectors.
Each document is represented as a count vector.

Count Vectors

Doc$_1$ = [0, 0, 1, 0, 0, 3…]

Doc$_2$ = [1, 0, 0, 0, 0, 2…]

Doc$_3$ = [0, 0, 0, 0, 0, 1…]

Doc$_4$ = [0, 0, 0, 0, 0, 0…]

Doc$_5$ = [1, 1, 0, 0, 0, 2…]

Doc$_6$ = [0, 0, 0, 0, 0, 1…]

Doc$_7$ = [1, 0, 0, 2, 0, 2…]

Doc$_8$ = [2, 0, 0, 0, 0, 0…]

Numbers represent the frequency of unique words in a document

Document-Term Matrix

A collection of count vectors is called the document-term matrix, or *dtm*.

Each row of the matrix is a different document.

Each column is a term.

Contains mostly zeroes (sparse).

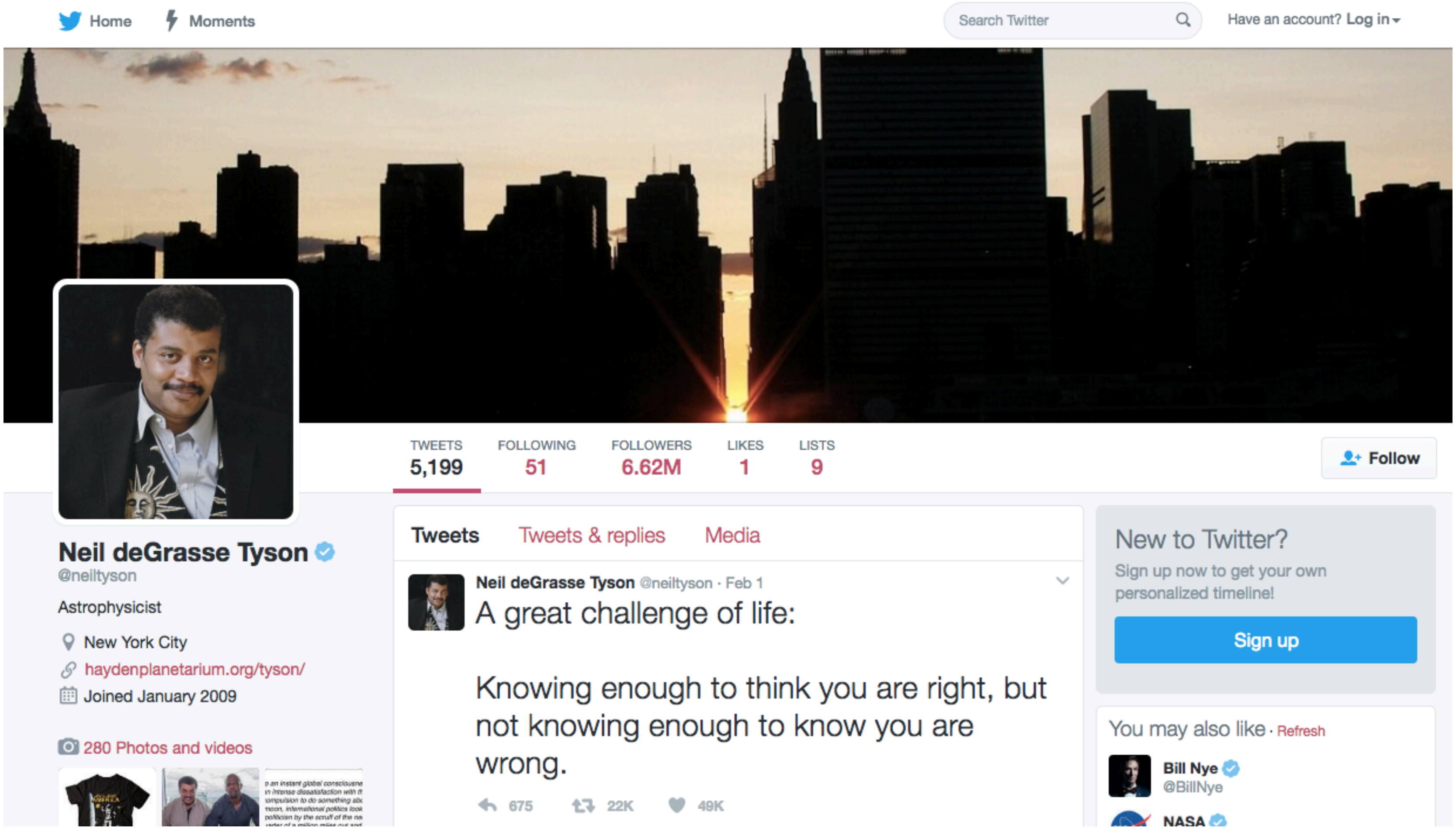
Document-Term Matrix

1	0	0	0	0	1	0	2
0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0
0	0	1	0	0	0	1	0
1	0	0	0	0	0	0	2
0	0	0	0	0	0	0	0
0	2	1	0	0	3	0	0

A real world example

The following slides show what this process actually looks like from start to finish.

Neil deGrasse Tyson’s Twitter



TWEETS	FOLLOWING	FOLLOWERS	LIKES	LISTS
5,199	51	6.62M	1	9

Neil deGrasse Tyson
@neiltyson

Astrophysicist

New York City
haydenplanetarium.org/tyson/
Joined January 2009

280 Photos and videos

Tweets   Tweets & replies   Media

Neil deGrasse Tyson @neiltyson · Feb 1

A great challenge of life:

Knowing enough to think you are right, but not knowing enough to know you are wrong.

675   22K   49K

New to Twitter?

Sign up now to get your own personalized timeline!

You may also like · Refresh

Bill Nye @BillNye

NASA

The Corpus of Tweets

A Corpus of Neil De Grasse Tyson's Tweets

Document	Text
1	A great challenge of life: Knowing enough to think you are right, but not knowing enough to know you are wrong.
2	I dream of a world where the truth is what shapes people's politics, rather than politics shaping what people think is true.
3	Let's Make America Smart Again.
4	How to Make America Great Again: Invest heavily in Science & Technology, the engines of tomorrow's growth economies.
5	If you wished upon that first Star you saw tonight in twilight, then it will not likely come true. You wished on planet Venus.
6	If ComicCon people ruled the world, international conflicts would be resolved entirely by plastic light saber fights in bars
7	When facts are what people want to be true, in spite of contrary evidence, witness the beginning of the end of an informed Democracy.

The Pre-Processed Corpus

Tweets after removing punctuation and stopwords, stemming

Document	Pre-processed text
1	great, challenge, life, know, enough, think, right, know, wrong
2	dream, world, true, shape, people, politic, rather, think, true, politic
3	america, smart, again
4	america, great, again, invest, heavily, science, technology, engine, tomorrow growth econom
5	wish, upon, first, star, tonight, twilight, likely, true, planet, venus
6	comiccon, people, rule, world, international, conflict, resolve, entire, plastic, light, saber, fight, bar
7	fact, people, true, spite, contrary, evidence, witness, beginning, end, informed democracy

The Document-Term Matrix

Document	**Terms**							
	great	america	true	people	think	science	know	politic
1	1	0	0	0	1	0	2	0
2	0	0	1	1	1	0	0	2
3	0	1	0	0	0	0	0	0
4	1	1	0	0	0	1	0	0
5	0	0	1	0	0	0	0	0
6	0	0	0	1	0	0	0	0
7	0	0	1	1	0	0	0	0

Why do we do this?

Creating a document-term matrix simplifies analysis by narrowing our focus to the ***core content*** of each document.

Why do we do this?

Core content varies from project to project.

Sometimes we care about common terms, but other times we care about rare terms.

These decisions affect the pre-processing stage, which words to keep or discard.

Why do we do this?

Regardless of the task, the process of converting words to numbers, documents to vectors, and corpora to matrices makes qualitative text machine readable.

Why do we do this?

After processing text, computers can read and synthesize content we are interested in to find broad patterns across thousands of documents.

Why do we do this?

Computers are basically doing what humans do – read texts – but on a larger scale and in a more systematic way.

Methods

Many NLP applications rely on machine learning algorithms, which take raw data as input and automatically detect patterns for output.

Machine learning methods may be *supervised* or *unsupervised*.

Supervised vs. Unsupervised

Supervised and unsupervised methods are often cast as competitors. But they have different objectives.

Supervised methods: you have predetermined categories and documents that need to be placed in those categories.

Unsupervised methods: you don't have a predetermined categorization scheme.

Unsupervised Methods

With unsupervised methods, we use modeling assumptions and underlying properties of texts to estimate a set of categories and simultaneously assign documents to those categories.

Unsupervised Methods

Unsupervised methods can identify categories of text that are theoretically useful, but perhaps understudied or previously unknown.

Also known as *topic modeling*.

Topic Models

Statistically speaking, a topic model is a probability mass function over words.

Substantively speaking, topics are distinct concepts.

How does Topic Modeling work?

Say we have a corpus of documents.

We only see the raw text.

But we do not know a priori the topics contained within these documents.

Topics are *hidden variables* that make up the underlying thematic structure of the corpora.

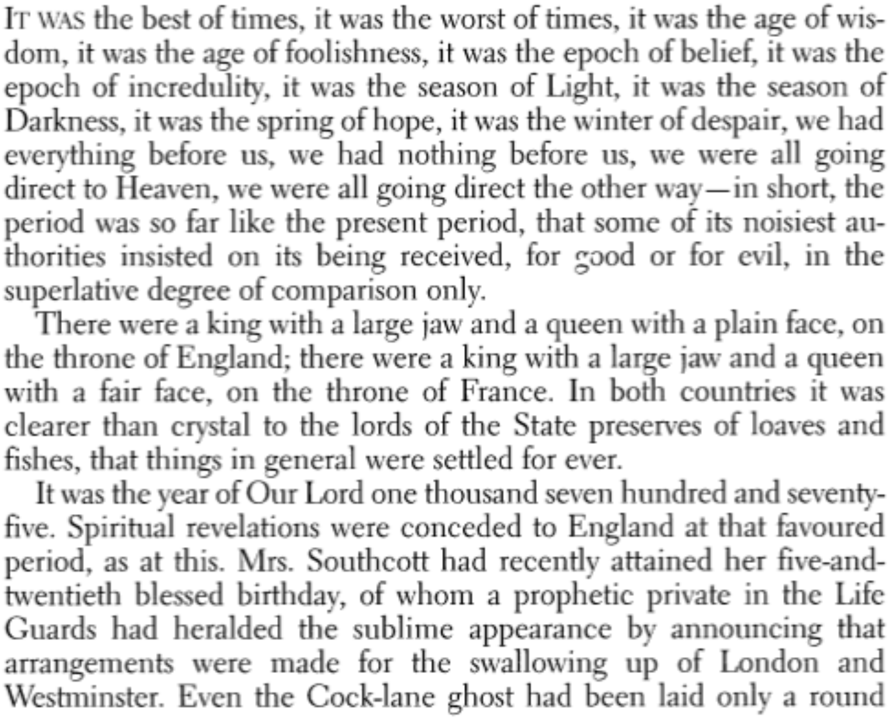
How does Topic Modeling work?

We can infer topics by computing their posterior distribution, conditioning on the documents themselves.

By using probability models, we can assign topic probabilities to unique words in each document.

From Start…

The raw document



IT WAS the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.

There were a king with a large jaw and a queen with a plain face, on the throne of England; there were a king with a large jaw and a queen with a fair face, on the throne of France. In both countries it was clearer than crystal to the lords of the State preserves of loaves and fishes, that things in general were settled for ever.

It was the year of Our Lord one thousand seven hundred and seventy-five. Spiritual revelations were conceded to England at that favoured period, as at this. Mrs. Southcott had recently attained her five-and-twentieth blessed birthday, of whom a prophetic private in the Life Guards had heralded the sublime appearance by announcing that arrangements were made for the swallowing up of London and Westminster. Even the Cock-lane ghost had been laid only a round

…to Finish

Topics

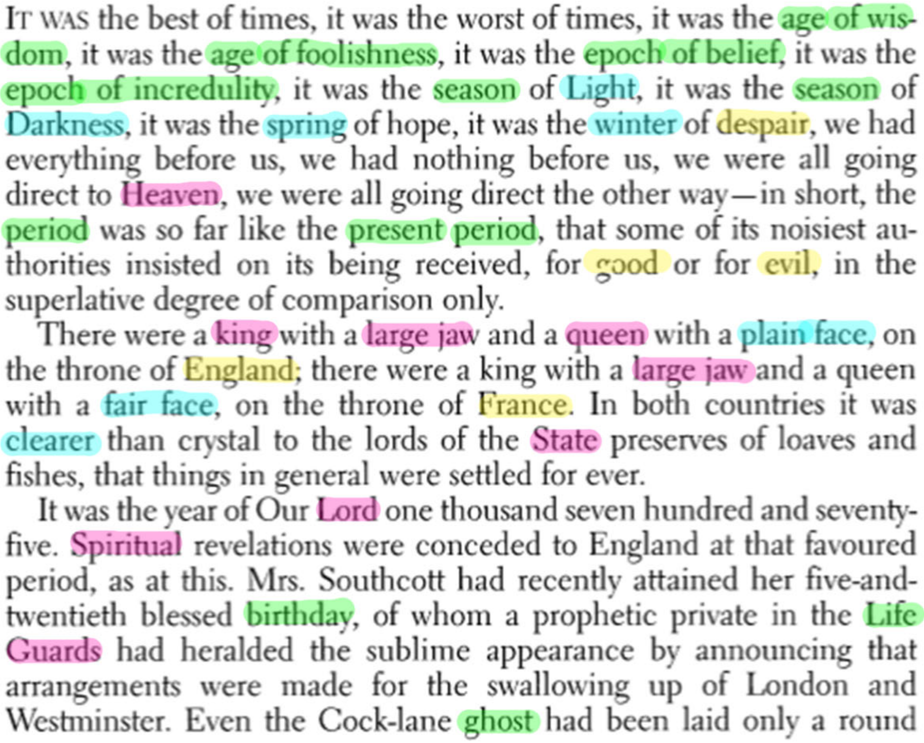
age	8.04
epoch	8.02
season	8.00
…	

light	8.01
dark	8.03
spring	7.09
…	

heaven	4.01
king	8.09
queen	9.01
…	

despair	8.00
evil	9.00
good	8.93
…	

Document



IT WAS the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.

There were a king with a large jaw and a queen with a plain face, on the throne of England; there were a king with a large jaw and a queen with a fair face, on the throne of France. In both countries it was clearer than crystal to the lords of the State preserves of loaves and fishes, that things in general were settled for ever.

It was the year of Our Lord one thousand seven hundred and seventy-five. Spiritual revelations were conceded to England at that favoured period, as at this. Mrs. Southcott had recently attained her five-and-twentieth blessed birthday, of whom a prophetic private in the Life Guards had heralded the sublime appearance by announcing that arrangements were made for the swallowing up of London and Westminster. Even the Cock-lane ghost had been laid only a round

Topic Models: a hierarchy

Topic models have a hierarchical structure:

The top level contains information about key features contained in the documents. This info is used to estimate the different topics.

Topic Models: a hierarchy

Topic models have a hierarchical structure:

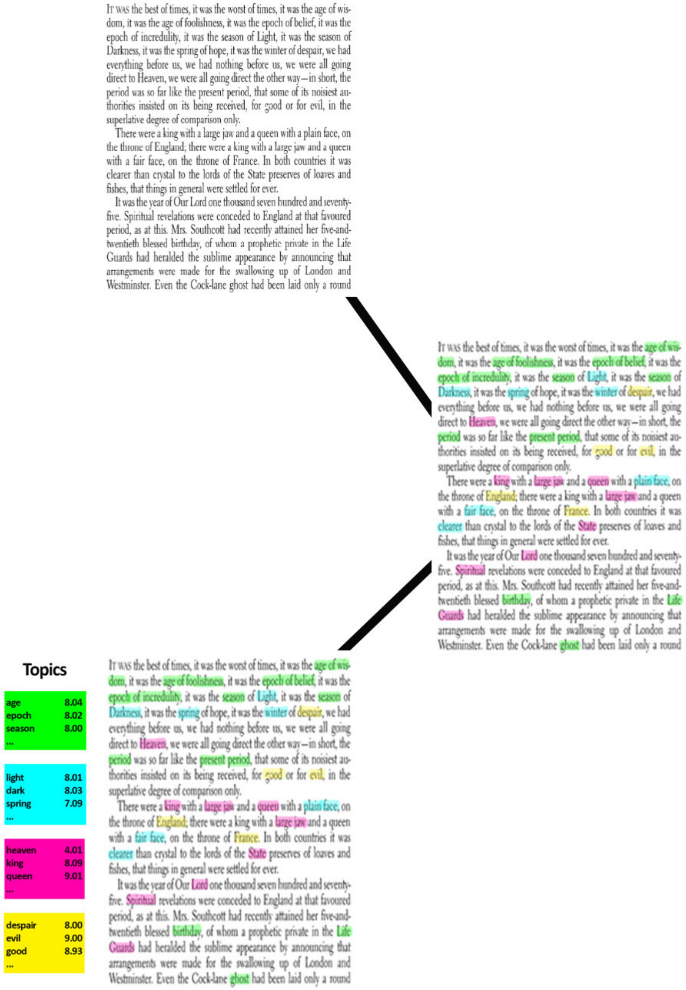
The middle level measures the extent to which each document is focused on the estimated topics.

Topic Models: a hierarchy

Topic models have a hierarchical structure:

- The bottom level assigns each document to a single topic.

Topic Models: a hierarchy



IT WAS the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.

There were a king with a large jaw and a queen with a plain face, on the throne of England; there were a king with a large jaw and a queen with a fair face, on the throne of France. In both countries it was clearer than crystal to the lords of the State preserves of loaves and fishes, that things in general were settled for ever.

It was the year of Our Lord one thousand seven hundred and seventy-five. Spiritual revelations were conceded to England at that favoured period, as at this. Mrs. Southcott had recently attained her five-and-twentieth blessed birthday, of whom a prophetic private in the Life Guards had heralded the sublime appearance by announcing that arrangements were made for the swallowing up of London and Westminster. Even the Cock-lane ghost had been laid only a round

IT WAS the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.

There were a king with a large jaw and a queen with a plain face, on the throne of England; there were a king with a large jaw and a queen with a fair face, on the throne of France. In both countries it was clearer than crystal to the lords of the State preserves of loaves and fishes, that things in general were settled for ever.

It was the year of Our Lord one thousand seven hundred and seventy-five. Spiritual revelations were conceded to England at that favoured period, as at this. Mrs. Southcott had recently attained her five-and-twentieth blessed birthday, of whom a prophetic private in the Life Guards had heralded the sublime appearance by announcing that arrangements were made for the swallowing up of London and Westminster. Even the Cock-lane ghost had been laid only a round

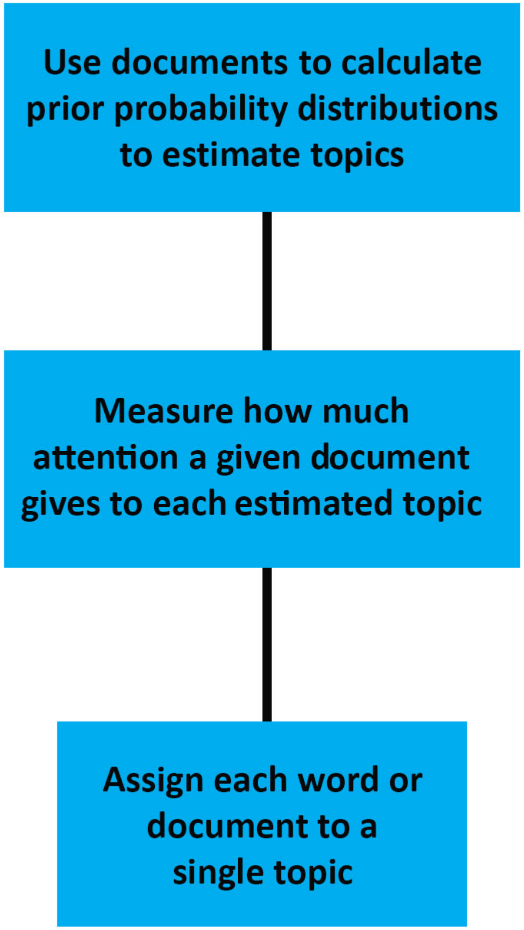
IT WAS the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.

There were a king with a large jaw and a queen with a plain face, on the throne of England; there were a king with a large jaw and a queen with a fair face, on the throne of France. In both countries it was clearer than crystal to the lords of the State preserves of loaves and fishes, that things in general were settled for ever.

It was the year of Our Lord one thousand seven hundred and seventy-five. Spiritual revelations were conceded to England at that favoured period, as at this. Mrs. Southcott had recently attained her five-and-twentieth blessed birthday, of whom a prophetic private in the Life Guards had heralded the sublime appearance by announcing that arrangements were made for the swallowing up of London and Westminster. Even the Cock-lane ghost had been laid only a round

Topics

age	8.04
epoch	8.02
season	8.00
...	
light	8.01
dark	8.03
spring	7.09
...	
heaven	4.01
king	8.09
queen	9.01
...	
despair	8.00
evil	9.00
good	8.93
...	



Use documents to calculate prior probability distributions to estimate topics

Measure how much attention a given document gives to each estimated topic

Assign each word or document to a single topic

Topic Models

Different types of probability models:

- Binomial
- Multinomial
- Latent Dirichlet Allocation

Topic Models

“Machine Learning Language Toolkit” (MALLET) is a topic modeling software that uses Latent Dirichlet Allocation. It is written in Java and executed in the terminal, but the output can be easily interpreted in Python.

Topic Models

Let’s try it out!