NLP 101

Gaurav Shahane

MIM, UMD

gshahane@umd.edu

Natural Language Processing (NLP)

- NLP is an application of Computational Linguistics: the study of processing languages with the application of computation.

- Applications:
  - Social media insights
  - Text logs of call center
  - Lots of documentation in Corporate, Research, and etc.
  - Siri, Cortana, Alexa, Viv (?)
  - Search Engines
  - Machine Translation

- Natural Language Toolkit (NLTK) is a Python package for NLP

Terminology

- Token: A sequence of characters
- Tokenizing: Splitting a text into Sentences or Words
- Stop Words: Unimportant words for language processing (such as a, an, the, am, is, are, you, he, she ..)
- Parsing: Breaking a text into parts as per some rules
- Stemming: Removing the suffixes from a word
- Lemmatization: Getting the base form of the word
- Named Entity Recognition: Identifying Names of Places, Persons, and Organizations from the given text.

Text Wrangling & Cleaning

- Parsing
- Raw Text
- Tokenizing
  - Sentence Tokenizing
  - Word Tokenizing
- Punctuation Removal
- Stop Words Removal
- Stemming
- Lemmatizing

steering you in the wrong direction; the moment you are old enough to take the wheel, responsibility lies with you. What is more, I cannot criticise my parents for hoping that I would never experience poverty. They had been poor themselves, and I have since been poor, and I quite agree with them that it is not an ennobling experience. Poverty entails fear, and stress, and sometimes depression; it means a thousand petty humiliations and hardships. Climbing out of poverty by your own efforts, that is indeed something on which to pride yourself, but poverty itself is romanticised only by fools. What I feared most for myself at your age was not poverty, but failure.

Subscribe to the Daily Gazette

Sign up for daily emails with the latest Harvard news.

At your age, in spite of a distinct lack of motivation at university, where I had spent far too long in the coffee bar writing stories, and far too little time at lectures, I had a knack for passing examinations, and that, for years, had been the measure of success in my life and that of my peers.
I am not dull enough to suppose that because you are young, gifted and well-educated, you have never known hardship or heartbreak. Talent and intelligence never yet inoculated anyone against the caprice of the Fates, and I do not for a moment suppose that everyone here has enjoyed an existence of unruffled privilege and contentment.

Text Wrangling & Cleaning

- Parsing
- Raw Text
- Tokenizing
  - Sentence Tokenizing
  - Word Tokenizing
- Punctuation Removal
- Stop Words Removal
- Stemming
- Lemmatizing

```
ultimately, we all have to decide for ourselves what constitutes
failure, but the world is quite eager to give you a set of
criteria if you let it. so i think it fair to say that by any
conventional measure, a mere seven years after my graduation
day, i had failed on an epic scale. an exceptionally short-lived
marriage had imploded, and i was jobless, a lone parent, and as
poor as it is possible to be in modern britain, without being
homeless. the fears that my parents had had for me, and that i
had had for myself, had both come to pass, and by every usual
standard, i was the biggest failure i knew.
```

Text Wrangling & Cleaning

- Parsing
- Raw Text
- Tokenizing
  - Sentence Tokenizing
  - Word Tokenizing
- Punctuation Removal
- Stop Words Removal
- Stemming
- Lemmatizing

```
['ultimately, we all have to decide for ourselves what
constitutes failure, but the world is quite eager to give you a
set of criteria if you let it.', 'so i think it fair to say that
by any conventional measure, a mere seven years after my
graduation day, i had failed on an epic scale.', 'an
exceptionally short-lived marriage had imploded, and i was
jobless, a lone parent, and as poor as it is possible to be in
modern britain, without being homeless.', 'the fears that my
parents had had for me, and that i had had for myself, had both
come to pass, and by every usual standard, i was the biggest
failure i knew.']
```

Text Wrangling & Cleaning

- Parsing
- Raw Text
- Tokenizing
  - Sentence Tokenizing
  - Word Tokenizing
- Punctuation Removal
- Stop Words Removal
- Stemming
- Lemmatizing

```
['ultimately', ',', 'we', 'all', 'have', 'to', 'decide', 'for',
'ourselves', 'what', 'constitutes', 'failure', ',', 'but',
'the', 'world', 'is', 'quite', 'eager', 'to', 'give', 'you',
'a', 'set', 'of', 'criteria', 'if', 'you', 'let', 'it', '.',
'so', 'i', 'think', 'it', 'fair', 'to', 'say', 'that', 'by',
'any', 'conventional', 'measure', ',', 'a', 'mere', 'seven',
'years', 'after', 'my', 'graduation', 'day', ',', 'i', 'had',
'failed', 'on', 'an', 'epic', 'scale', '.', 'an',
'exceptionally', 'short', '-', 'lived', 'marriage', 'had',
'imploded', ',', 'and', 'i', 'was', 'jobless', ',', 'a', 'lone',
'parent', ',', 'and', 'as', 'poor', 'as', 'it', 'is',
'possible', 'to', 'be', 'in', 'modern', 'britain', ',',
'without', 'being', 'homeless', '.', 'the', 'fears', 'that',
'my', 'parents', 'had', 'had', 'for', 'me', ',', 'and', 'that',
'i', 'had', 'had', 'for', 'myself', ',', 'had', 'both', 'come',
'to', 'pass', ',', 'and', 'by', 'every', 'usual', 'standard',
',', 'i', 'was', 'the', 'biggest', 'failure', 'i', 'knew', '.']
```

Text Wrangling & Cleaning

- Parsing
- Raw Text
- Tokenizing
  - Sentence Tokenizing
  - Word Tokenizing
- Punctuation Removal
- Stop Words Removal
- Stemming
- Lemmatizing

```
['ultimately', 'we', 'all', 'have', 'to', 'decide', 'for',
'ourselves', 'what', 'constitutes', 'failure', 'but', 'the',
'world', 'is', 'quite', 'eager', 'to', 'give', 'you', 'a',
'set', 'of', 'criteria', 'if', 'you', 'let', 'it', 'so', 'i',
'think', 'it', 'fair', 'to', 'say', 'that', 'by', 'any',
'conventional', 'measure', 'a', 'mere', 'seven', 'years',
'after', 'my', 'graduation', 'day', 'i', 'had', 'failed', 'on',
'an', 'epic', 'scale', 'an', 'exceptionally', 'short', 'lived',
'marriage', 'had', 'imploded', 'and', 'i', 'was', 'jobless',
'a', 'lone', 'parent', 'and', 'as', 'poor', 'as', 'it', 'is',
'possible', 'to', 'be', 'in', 'modern', 'britain', 'without',
'being', 'homeless', 'the', 'fears', 'that', 'my', 'parents',
'had', 'had', 'for', 'me', 'and', 'that', 'i', 'had', 'had',
'for', 'myself', 'had', 'both', 'come', 'to', 'pass', 'and',
'by', 'every', 'usual', 'standard', 'i', 'was', 'the',
'biggest', 'failure', 'i', 'knew']
```

Text Wrangling & Cleaning

- Parsing
- Raw Text
- Tokenizing
  - Sentence Tokenizing
  - Word Tokenizing
- Punctuation Removal
- Stop Words Removal
- Stemming
- Lemmatizing

```
['ultimately', 'decide', 'constitutes', 'failure', 'world',
'quite', 'eager', 'give', 'set', 'criteria', 'let', 'think',
'fair', 'say', 'conventional', 'measure', 'mere', 'seven',
'years', 'graduation', 'day', 'failed', 'epic', 'scale',
'exceptionally', 'short', 'lived', 'marriage', 'imploded',
'jobless', 'lone', 'parent', 'poor', 'possible', 'modern',
'britain', 'without', 'homeless', 'fears', 'parents', 'come',
'pass', 'every', 'usual', 'standard', 'biggest', 'failure',
'knew']
```

Text Wrangling & Cleaning

- Parsing
- Raw Text
- Tokenizing
  - Sentence Tokenizing
  - Word Tokenizing
- Punctuation Removal
- Stop Words Removal
- Stemming
- Lemmatizing

```
['ultim', 'decid', 'constitut', 'failur', 'world', 'quit',
'eager', 'give', 'set', 'criteria', 'let', 'think', 'fair',
'say', 'convent', 'measur', 'mere', 'seven', 'year', 'graduat',
'day', 'fail', 'epic', 'scale', 'except', 'short', 'live',
'marriag', 'implod', 'jobless', 'lone', 'parent', 'poor',
'possibl', 'modern', 'britain', 'without', 'homeless', 'fear',
'parent', 'come', 'pass', 'everi', 'usual', 'standard',
'biggest', 'failur', 'knew']
```

Text Wrangling & Cleaning

- Parsing
- Raw Text
- Tokenizing
  - Sentence Tokenizing
  - Word Tokenizing
- Punctuation Removal
- Stop Words Removal
- Stemming
- Lemmatizing

```
['ultimately', 'decide', 'constitutes', 'failure', 'world',
'quite', 'eager', 'give', 'set', 'criterion', 'let', 'think',
'fair', 'say', 'conventional', 'measure', 'mere', 'seven',
'year', 'graduation', 'day', 'failed', 'epic', 'scale',
'exceptionally', 'short', 'lived', 'marriage', 'imploded',
'jobless', 'lone', 'parent', 'poor', 'possible', 'modern',
'britain', 'without', 'homeless', 'fear', 'parent', 'come',
'pas', 'every', 'usual', 'standard', 'biggest', 'failure',
'knew', 'passed', 'one']
```

pass[1]
/pas/

verb

1. move or cause to move in a specified direction.
   "he passed through towns and villages"
   synonyms: go, proceed, move, progress, make one's way, travel
   "the traffic passing through the village"

2. go past or across; leave behind or on one side in proceeding.
   "she passed a rest area with a pay phone"

Useful applications

- Parts-Of-Speech (POS) Tagging
- Named Entity Chunking
- Frequency Distribution

```
[('ultimately', 'RB'), (',', ','), ('we', 'PRP'), ('all', 'DT'),
('have', 'VBP'), ('to', 'TO'), ('decide', 'VB'), ('for', 'IN'),
('ourselves', 'PRP'), ('what', 'WP'), ('constitutes', 'VBZ'),
('failure', 'NN'), (',', ','), ('but', 'CC'), ('the', 'DT'),
('world', 'NN'), ('is', 'VBZ'), ('quite', 'RB'), ('eager',
'JJ'), ('to', 'TO'), ('give', 'VB'), ('you', 'PRP'), ('a',
'DT'), ('set', 'NN'), ('of', 'IN'), ('criteria', 'NNS'), ('if',
'IN'), ('you', 'PRP'), ('let', 'VBP'), ('it', 'PRP'), ('.',
'.'), ('so', 'RB'), ('i', 'JJ'), ('think', 'VBP'), ('it',
'PRP'), ('fair', 'VBZ'), ('to', 'TO'), ('say', 'VB'), ('that',
'IN'), ('by', 'IN'), ('any', 'DT'), ('conventional', 'JJ'),
('measure', 'NN'), (',', ','), ('a', 'DT'), ('mere', 'JJ'),
('seven', 'CD'), ('years', 'NNS'), ('after', 'IN'), ('my', 'PRP
$'), ('graduation', 'NN'), ('day', 'NN'), (',', ','), ('i',
'NN'), ('had', 'VBD'), ('failed', 'VBN'), ('on', 'IN'), ('an',
'DT'), ('epic', 'JJ'), ('scale', 'NN'), ('.', '.'), ('an',
'DT'), ('exceptionally', 'RB'), ('short', 'JJ'), ('-', ':'),
('lived', 'JJ'), ('marriage', 'NN'), ('had', 'VBD'),
```

Useful applications

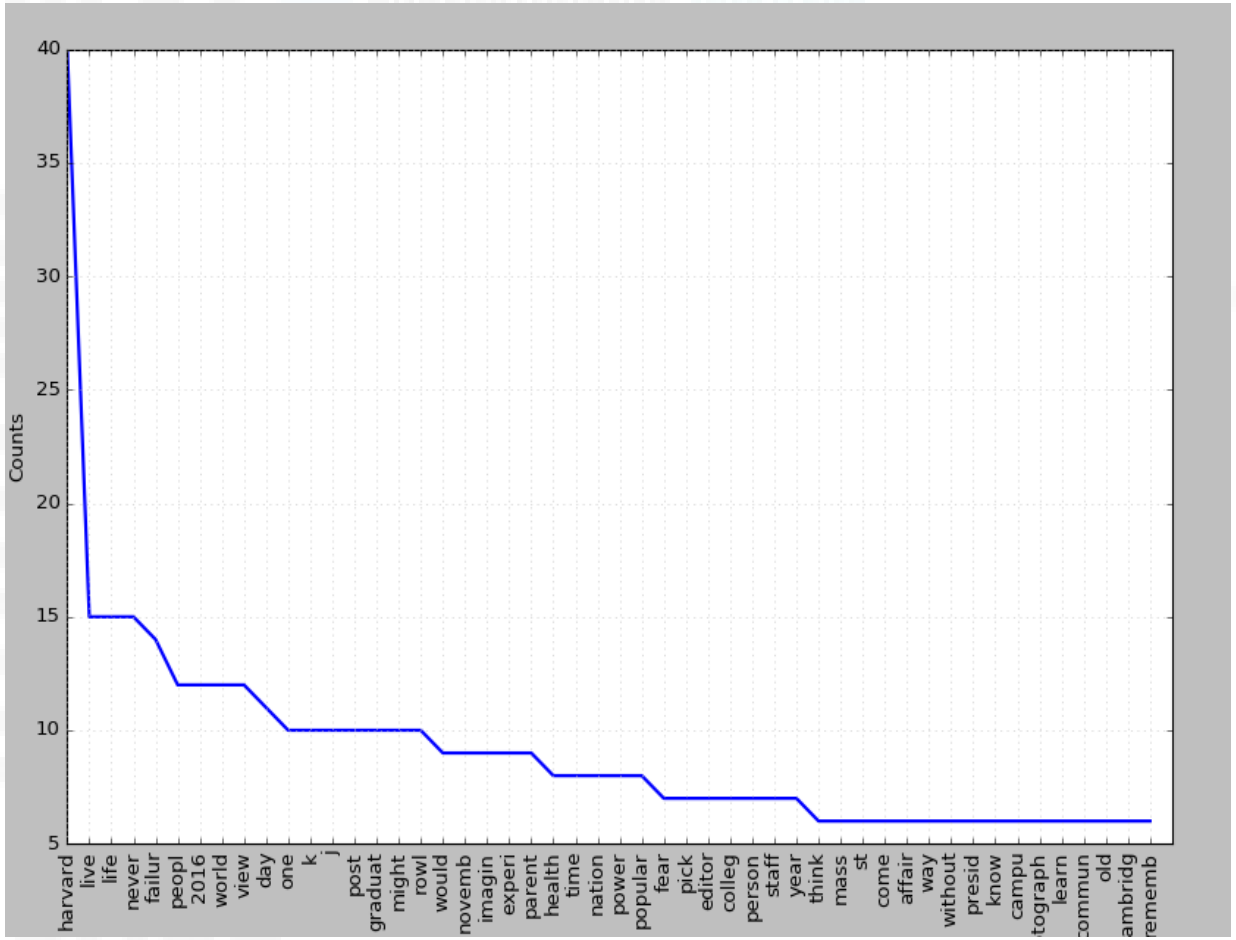
- Parts-Of-Speech (POS) Tagging
- Named Entity Chunking
- Frequency Distribution

```
exceptionally/RB
 short/JJ
 -/:
 lived/JJ
 marriage/NN
 had/VBD
 imploded/VBN
 ,/,
 and/CC
 I/PRP
 was/VBD
 jobless/JJ
 ,/,
 a/DT
 lone/NN
 parent/NN
 ,/,
 and/CC
 as/RB
 poor/JJ
 as/IN
 it/PRP
 is/VBZ
 possible/JJ
 to/TO
 be/VB
 in/IN
 modern/JJ
 (GPE Britain/NNP)
 ,/,
 without/IN
 being/VBG
 homeless/NN
 ./.
```

```
The/DT
  text/NN
  of/IN
  (PERSON Joanne/NNP Rowling/NNP)
  aka/NN
  J/NNP
  ./.
  K/NNP
  ./.
  Rowling/VBG
  ’/JJ
  s/JJ
  speech/NN
  at/IN
  (ORGANIZATION Harvard/NNP University/NNP)
  ,/,
  (GPE Cambridge/NNP)
  ,/,
  (ORGANIZATION MA/NNP))
```

Useful applications

- Parts-Of-Speech (POS) Tagging
- Named Entity Chunking
- Frequency Distribution



Corpus and Bag of Words

- Corpus is a simple list with a lot of text…

[text_1, text_2, text_3,……., text_n]

corpus = [

'UMD played **Duke** in basketball',

'Duke lost the basketball game',

]

- This text can belong to a newspaper agency, where each text could be an article, or the text can belong to a corporate where each text could be a meeting memo, email, call center notes, or any text data.

- Bag of Words represents set of words of a given text, or a corpus.

Text_1 = [**1** 1 **0** 1 0 1 0 1]

Text_2 = [**1** 1 **1** 0 1 0 1 0]

{u'duke': 1, u'basketball': 0**, u'lost':** 4, u'played': 5, u'game': 2,u‘umd': 7, u'in': 3, u'the': 6}

- This text can belong to a newspaper agency, where each text could be an article, or the text can belong to a corporate where each text could be a meeting memo, email, call center notes, or any text data.

Term Freq. – Inverse Document Freq. (TFIDF)

Based on the fact that a term (or token) will be unique to a document…

if it ***appears a lot in the given text,***

and ***not much in all the other documents in a corpus.***

Eg. ‘Failure’ will be unique to the JK Rowling speech, and might not be present in very other document on Harvard’s Article Corpus. So, it will have a greater score for the document.

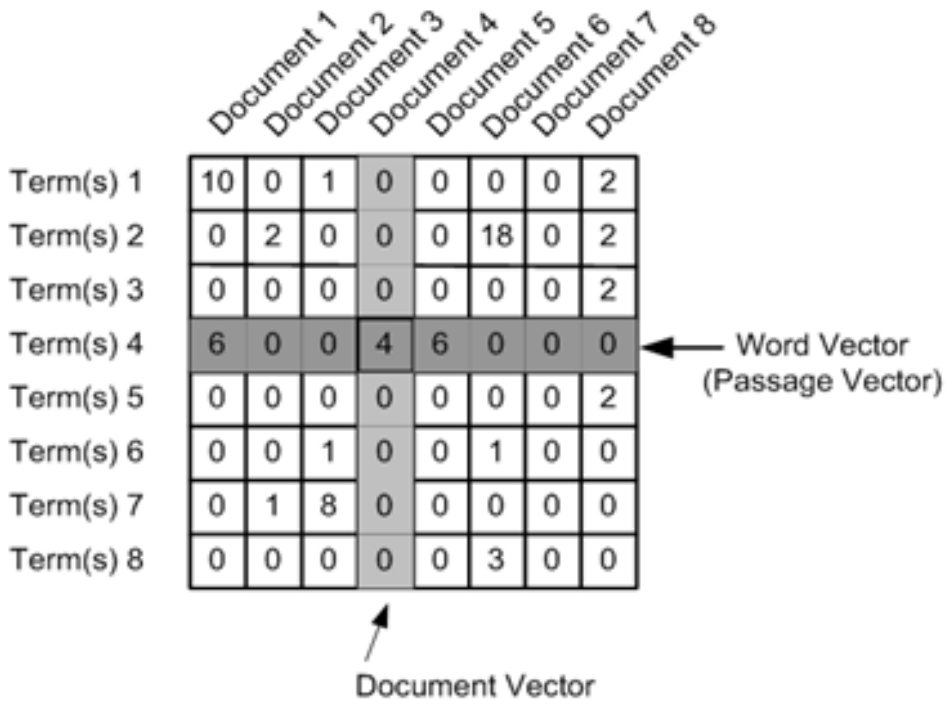


Image Source: http://brandonrose.org/clustering

NLP Applications

- Topic Modelling
- Information Summarization
- Machine Translation
- Information Retrieval
- Speech Recognition
- Text Classification
- Sentiment Analysis
- Information Extraction
- Language Detection

Resources

- The code and PPT: https://github.com/gaurav-shahane/Natural-Language-Processing
- For more details on NLP Practices: http://www.nltk.org/book/
- Scikit-Learn TFIDF: http://scikit-learn.org/stable/modules/feature_extraction.html
- Stanford NLP Group: http://nlp.stanford.edu/
- Other several NLTK and Machine Learning Books from O'Reilly Publication.


Thank You
Vinaka
köszönöm
Kiitos
Maake
Asante
Shukria
Dhanyavadagalu
Manana
Dankon
Matondo
감사합니다
Kam Sah Hammida
Dank Je
Dankscheen
спасибо
Mauruuru
Biyan
Blagodaram
Ngiyabonga
Dziekuje
Chokrane
Diolch i Chi
Terima Kasih
Taiku
Tack
Juspaxar
Arigato
Grazie
Gracias
Mochchakkeram
நன்றி
Bedankt
Dakujem
धन्यवाद
cảm ơn bạn
Tingki
Paldies
Ua Tsaug Rau Koj
Grazas
Gratias Tibi
Obrigado
Děkuji
Nirringrazzjak
Di Ou Mési
Kia Ora
Kop Khun Khap
ありがとう
Suksama
Hvala
Welalin
Djiere Dieuf
Rahmat
Nuwun
Matur
谢
谢
Misaotra
Danke
Merci
Go Raibh Maith Agat
Eskerrik Asko
ХВала
Salamat
ขอขอบคุณคุณ
Najis Tuke