
How would you like to make more money when applying for grant funding?

Text-Mining on EPSRC grant abstracts

Sai Govindarajan 956299

Aims and Motivations

Aims

- Perform text analysis
- Output insightful information

Motivations

- Earn more money
- Reduce human error

Presentation: Overview

- Method
- Language & Library
- Current Stage

Method

Text-Mining

Making web-page and beautiful soup

Latent Dirichlet Allocation

Text-Mining?

- Deriving meaningful information from given text
- Consists of numerous stages:

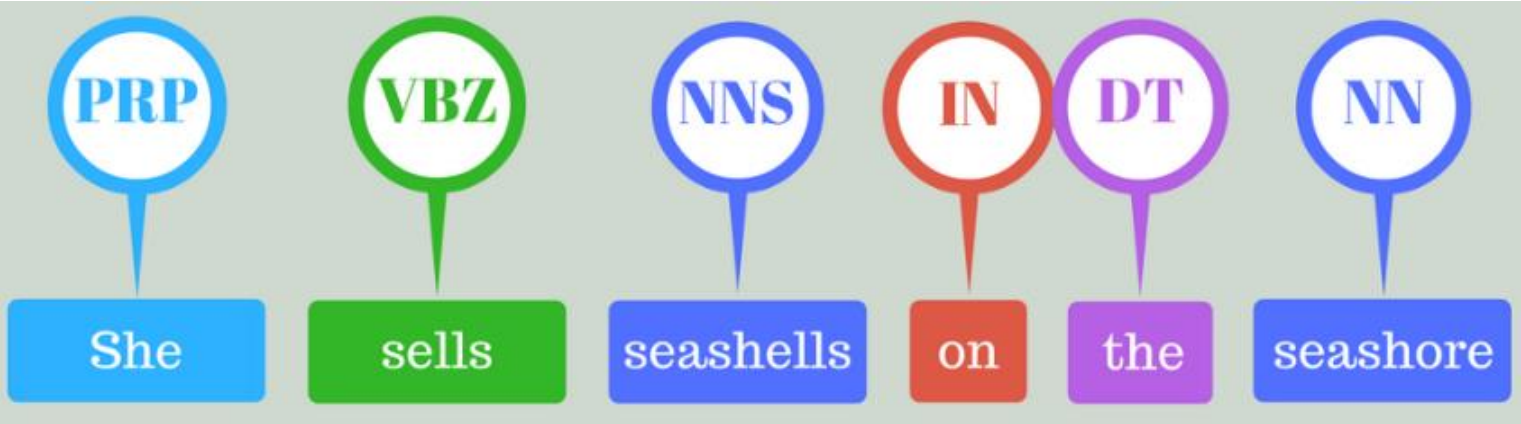
Tokenization

[We are a family]= [‘We’, ’are’, ’a’, ’family’]

Stemming & Lemmatization

Stemming
‘Running’ = ‘Run’
‘Better’ = ‘Bette’/ ’Bett’

Stemming & Lemmatization
‘Better’ = ‘Good’



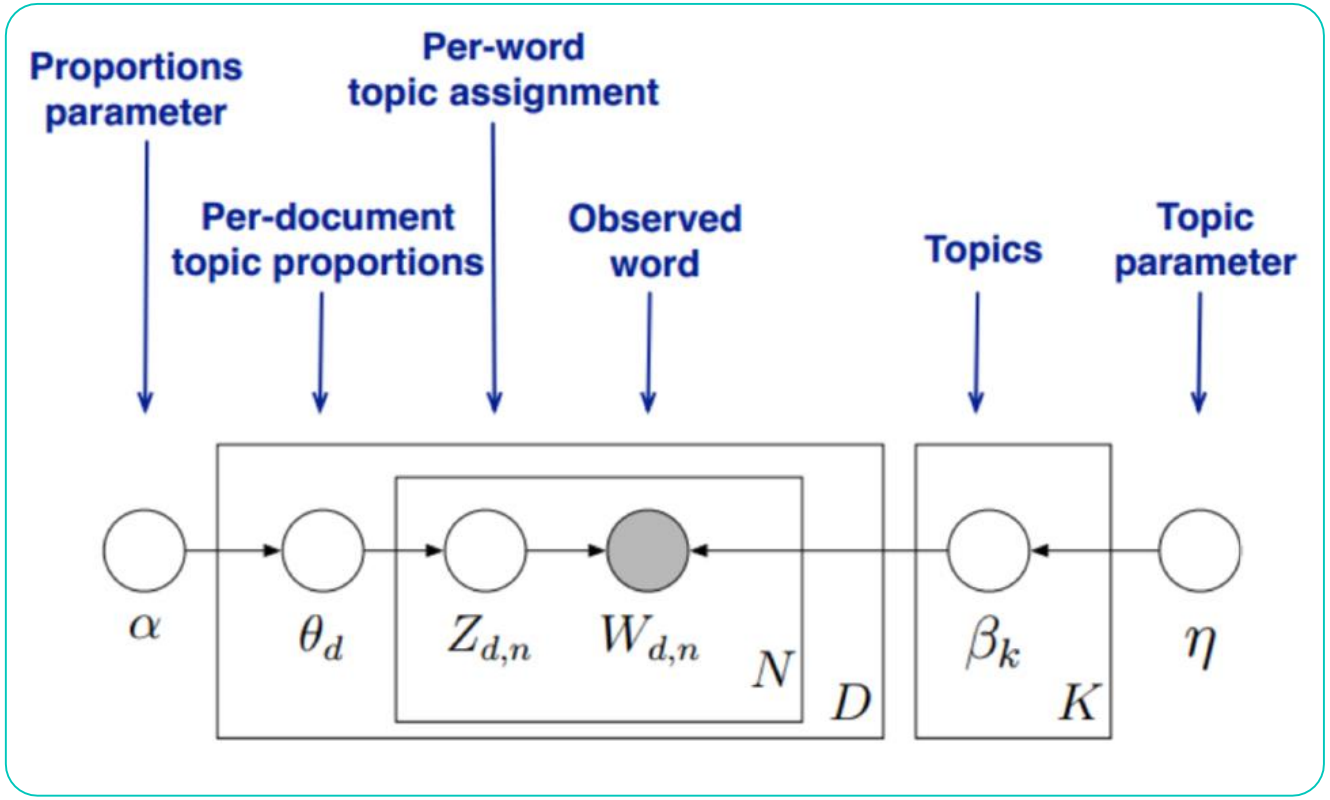
Part Of Speech Tagging

Named Entity Recognition

Chunking

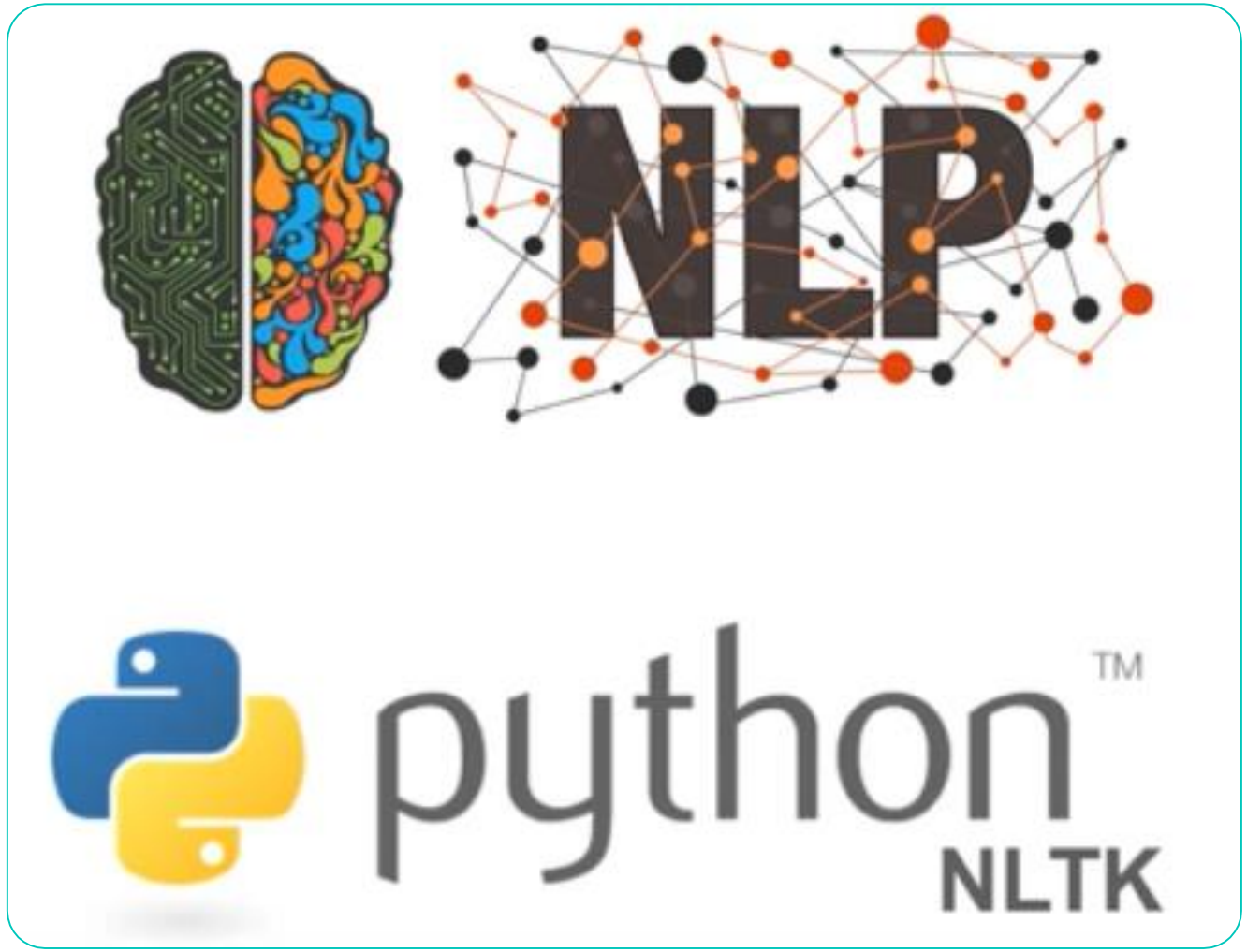
Making the website

Latent Dirichlet Allocation



Language & Library

- Language
  - Python
- Library
  - Natural Language Tool Kit (NLTK)
  - Beautiful Soup for information retrieval
  - Pandas for topic modelling



Steps done so far...

Sample Text

```
hey say too few people now carry the gene for blondes to last beyond the next
blonde hair is caused by a recessive gene . In order for a child to have blond
have blonde hair , it must have the gene on both sides of the family in the g
ere is a disadvantage of having that gene or by chance . They do n't disappear
des would disappear is if having the gene was a disadvantage and I do not thin
```

Tokenization

```python
import nltk
import nltk.corpus
from nltk.tokenize import sent_tokenize,word_tokenize

file = open("text.txt","r")
content = file.read()

tokens = nltk.word_tokenize(content)
print(sent_tokenize(content))
print(tokens)
```

```
['hey say too few people now carry the gene for blondes to last beyond the next\nblonde hair is caused by a recessive gene .', 'In order for a child to have blond\nhave blonde hair , it must have t
['hey', 'say', 'too', 'few', 'people', 'now', 'carry', 'the', 'gene', 'for', 'blondes', 'to', 'last', 'beyond', 'the', 'next', 'blonde', 'hair', 'is', 'caused', 'by', 'a', 'recessive', 'gene', '.',
```

```
import nltk
import nltk.corpus

from nltk.tokenize import *
from nltk.stem.snowball import SnowballStemmer
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer

from nltk.corpus import brown
```

```
['hey say too few people now carry the gene for blondes to last beyond the next\nblonde hair is caused by a recessive gene .', 'In order for a child to have blond\nhave blonde hair , it must have t
['hey', 'say', 'too', 'few', 'people', 'now', 'carry', 'the', 'gene', 'for', 'blondes', 'to', 'last', 'beyond', 'the', 'next', 'blonde', 'hair', 'is', 'caused', 'by', 'a', 'recessive', 'gene', '.',
hey say too few peopl now carri the gene for blond to last beyond the next blond hair is caus by a recess gene . in order for a child to have blond have blond hair , it must have the gene on both s
hey say too few people now carry the gene for blondes to last beyond the next
blonde hair is caused by a recessive gene . In order for a child to have blond
have blonde hair , it must have the gene on both sides of the family in the g
ere is a disadvantage of having that gene or by chance . They do n't disappear
des would disappear is if having the gene was a disadvantage and I do not thin
```

Stemming & Lemmatization

```
file = open("text.txt","r")
content = file.read()

tokens = nltk.word_tokenize(content)

stemmer1 = SnowballStemmer('english')
stemmer2 = PorterStemmer()
lemma = WordNetLemmatizer()

def stemmer(content):
    token_words=word_tokenize(content)
    token_words
    stem = []
    for word in token_words:
        stem.append(stemmer1.stem(word))
        stem.append(" ")
    return "".join(stem)


output = stemmer(content)
lemmatized_output = ' '.join([lemma.lemmatize(word) for word in content])

print(sent_tokenize(content))
print(tokens)
print(output)
print(lemmatized_output)
```

Steps to be done

- Text Mining
  - Part Of Speech or POS Tagging
  - Named entity recognition
  - Chunking
- Making the web page
- Latent Dirichlet Allocation

To Conclude

- Keypoints
  - Aims- Perform Text Analysis & Derive insightful information
  - Motivation – Money & Reducing human Error
  - Method – Text-Mining, Making Website, Latent Dirichlet Allocation
  - Language & Libraries – Python & NLTK, Beautiful Soup, Pandas