Pycon17

Unstructured Text Analysis

Introduction

- Manufacturing Engineer from GIKI
- Technology enthusiast and a Data Science Consultant at IBM
- Python Evangelist
- Transition from Manufacturing to Data Science.

Agenda

- A refresher on building classification models with sckit-learn
- Representing text as numerical data using Vector space models
- Reading a text-based dataset into pandas
- Data Preprocessing
- Vectorizing our dataset
- Building and evaluating a model
- Comparing models
- Further Improvements

Refresher

"**Features**" are also known as predictors, inputs, or attributes. The "**response**" is also known as the target, label, or output.

"**Observations**" are also known as samples, instances, or records.

In order to **build a model**, the features must be **numeric**, and every observation must have the **same features in the same order**.

Representing text as numerical data

From the scikit-learn documentation:

•Text Analysis is a major application field for machine learning algorithms. However the raw data, a sequence of symbols cannot be fed directly to the algorithms themselves as most of them expect **numerical feature vectors with a fixed size** rather than the **raw text documents with variable length**

Data Frame

From tutorial points documentation:

•A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns.

Features of DataFrame:

•Columns can be of different types

•Size – Mutable

•Can Perform Arithmetic operations on rows and columns

Pre Processing-I

Stop Words

•Stop words are a set of commonly used words in any language.

•Stop words are commonly eliminated from many text processing applications because these words can be distracting, non-informative (or non-discriminative) and are additional memory overhead

Word Replacement

•In real data there are a lot of data in which there are a lot of spelling mistakes

•The Program doesn't know about all this so we have to replace them correctly in order to avoid noise

Pre Processing-II

- **Stemming and Lemmatization**

Stemming is the process of removing and replacing word suffixes to arrive at a common root for of the word. Lemmas differ from stems in that a lemma is a canonical form of the word, while a stem may not be a real word.

For example, from "produced", the lemma is "produce", but the stem is "produc-". This is because there are words such as production.

- **Template Extraction**

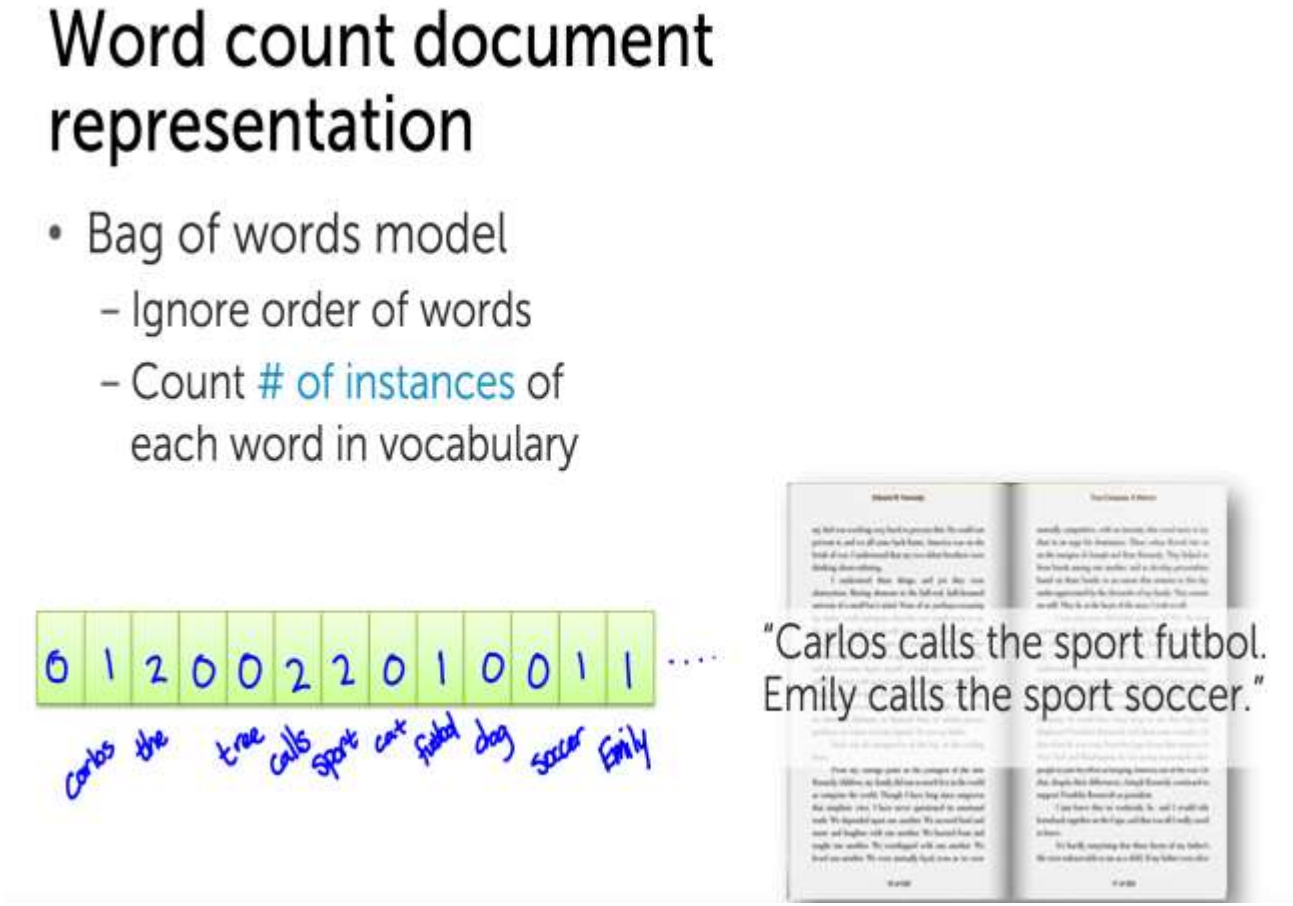
There are some particular patterns in our data which are useful while all others are just the noise.
So in order to improve our result we can remove it before applying our algorithm.

Vector Space Model- COUNT VECTORIZER

- **Count Vectorizer**

Transform a collection of text samples to a vector of token counts.

- minDF (float) - ignore tokens that have a samples frequency strictly lower than the given threshold. This value is also called cut-off in the literature.
- **Tokenizers**
- WhitespaceTokenizer - select tokens by whitespace.
- WordTokenizer - select tokens of 2 or more alphanumeric characters (punctuation is completely ignored and always treated as a token separator).



Issues with word counts – Doc length

1	0	0	0	5	3	0	0	1	0	0	0	0

2	0	0	0	10	6	0	0	2	0	0	0	0

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

Similarity = 13

6	0	0	0	4	0	0	2	0	2	0	0	0

Similarity = 52

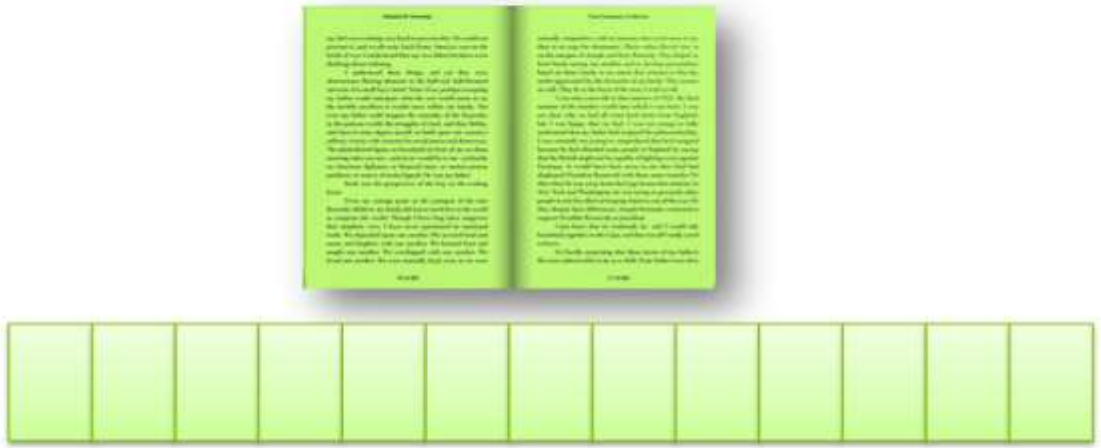
Solution = normalize

1	0	0	0	5	3	0	0	1	0	0	0	0

$$\sqrt{(1^2 + 5^2 + 3^2 + 1^2)}$$

1/6	0	0	0	5/6	3/6	0	0	1/6	0	0	0	0

Issues with word counts – Rare words



Common words in doc: “the”, “player”, “field”, “goal”

Dominate rare words like: “futbol”, “Messi”

Important words

- Do we want only rare words to dominate???
- What characterizes an important word?
  - Appears frequently in document (common locally)
  - Appears rarely in corpus (rare globally)
- Trade off between local frequency and global rarity

Vector Space Model-- TF-idf

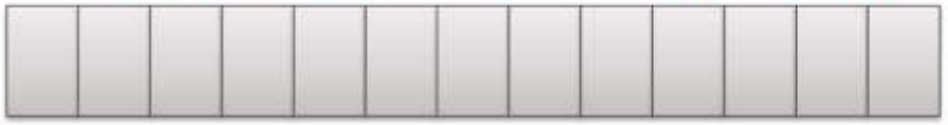
- Tf-idf stands for *term frequency-inverse document frequency*, and the tf-idf weight is a weight often used in information retrieval and text mining.
- This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus.
- The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus
- Tf-idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification.

For further information visit : http://www.tfidf.com/

TF-IDF document representation

- Term frequency – inverse document frequency (tf-idf)
- Term frequency

- Inverse document frequency



$$\log \frac{\#\ \text{docs}}{1 + \#\ \text{docs using word}}$$

Building and evaluating a model

WE WILL BE USING 2 DIFFERENT METHODS TO IDENTIFY THE DIFFERENCES AND EVALUATE THE RESULT(1 is Probabilistic and 2nd is Linear)

Multinomial Naive Bayes:

The multinomial Naive Bayes classifier is suitable for classification with **discrete features** (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work.

Logistic regression:

Logistic regression, despite its name, is a **linear model for classification** rather than regression. Logistic regression is also known in the literature as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function.

FURTHER IMPROVMENTS

- DEEPER DATA UNDERSTANDING
- PARAMETER TUNING
- CROSS VALIDATION : http://scikit-learn.org/stable/modules/cross_validation.html
- FEATURE ENGINEERING e.g WORD2VEC : https://elitedatascience.com/feature-engineering
- Algorithm Selection : https://elitedatascience.com/algorithm-selection

Contact Details

Name : Muhammad Usama Islam

Contact : +923400008836

Email Address : usamayawar@gmail.com

Linkedin Url : https://www.linkedin.com/in/muhammad-usama-islam-9b46ba96/

THANKYOU FOR YOUR TIME