

Lab Exercise Report on

WORKING OF BAG-OF-WORDS MODEL

Adikar Bharath N V S (171IT202)

Under the Guidance of,

Dr. Sowmya Kamath S.

Department of Information Technology, NITK Surathkal

Date of Submission: 15th October 2020

in partial fulfillment for the award of the degree

of

Bachelor of Technology

In

Information Technology

At



Department of Information Technology

National Institute of Technology Karnataka, Surathkal

October 2020

Table of Contents

1	Results	1
1.1	Introduction	1
1.2	Preprocessing	1
1.2.1	Stopword removal	1
1.2.2	Tokenization	2
1.2.3	Stemming	2
1.2.4	Lemmatization	3
1.3	Vocabulary Construction	3
1.4	N-gram modelling	3
1.5	Text Corpora	4
1.5.1	Gutenberg Corpus	4
1.5.2	Brown Corpus	4

Chapter 1

Results

1.1 Introduction

Whenever an algorithm in NLP is applied, it works on a collection of numbers. Thus, we cannot feed a set of words, paragraphs, text documents etc. to the algorithm as input directly, and must be converted into a representation in the form of numbers. Hence, the Bag-of-Words model is used to convert any text into its namesake, which keeps a count of the total occurrences of the most frequently used words.

The working of the model can be visualized using a table, with a count of each word displayed.

Before displaying the working of the model, some preprocessing of the text is required.

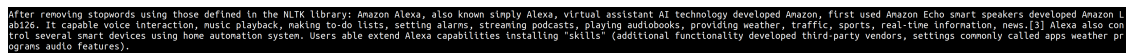
1.2 Preprocessing

1.2.1 Stopword removal

Stopwords are words in a language that do not add any meaning or semantic information to a sentence. These can be removed while keeping the meaning of the sentence intact. For example, the English words

- and
- the
- but

are stopwords. Since the selection of stopwords is highly subjective, the 'stopwords' package in the nltk library is downloaded for use and applied on a paragraph as demonstrated below:



After removing stopwords using those defined in the NLTK library: Amazon Alexa, also known simply as Alexa, virtual assistant AI technology developed Amazon, first used Amazon Echo smart speakers developed Amazon Lab126. It is capable of voice interaction, music playback, making to-do lists, setting alarms, streaming podcasts, playing audiobooks, and providing weather, traffic, sports, and other real-time information, such as news. [3] Alexa can also control several smart devices using itself as a home automation system. Users are able to extend the Alexa capabilities by installing "skills" (additional functionality developed by third-party vendors, in other settings more commonly called apps such as weather programs and audio features).

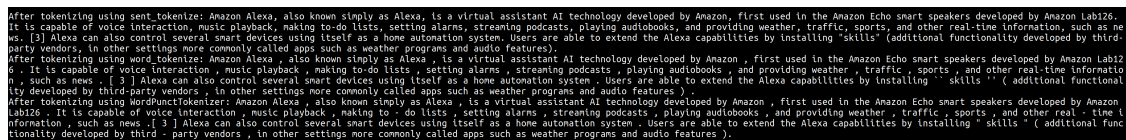
Figure 1.1: Stopword removal

1.2.2 Tokenization

Tokenization refers to the splitting up of words in a text into smaller words (for English), or even the creation of new words (in other languages). The NLTK library offers several ways of tokenization, such as:

- Sentence Tokenization: Splits text into sentences
- Word Tokenization: Splits text into words
- WordPunctuationTokenizer: Splits text into words as well as punctuation marks

Their working is demonstrated below:



After tokenizing using sent_tokenize: Amazon Alexa, also known simply as Alexa, is a virtual assistant AI technology developed by Amazon, first used in the Amazon Echo smart speakers developed by Amazon Lab126. It is capable of voice interaction, music playback, making to-do lists, setting alarms, streaming podcasts, playing audiobooks, and providing weather, traffic, sports, and other real-time information, such as news. [3] Alexa can also control several smart devices using itself as a home automation system. Users are able to extend the Alexa capabilities by installing "skills" (additional functionality developed by third-party vendors, in other settings more commonly called apps such as weather programs and audio features).

After tokenizing using word_tokenize: Amazon Alexa, also known simply as Alexa, is a virtual assistant AI technology developed by Amazon, first used in the Amazon Echo smart speakers developed by Amazon Lab126. It is capable of voice interaction, music playback, making to-do lists, setting alarms, streaming podcasts, playing audiobooks, and providing weather, traffic, sports, and other real-time information, such as news. [3] Alexa can also control several smart devices using itself as a home automation system. Users are able to extend the Alexa capabilities by installing "skills" (additional functionality developed by third-party vendors, in other settings more commonly called apps such as weather programs and audio features).

After tokenizing using WordPunctTokenizer: Amazon Alexa, also known simply as Alexa, is a virtual assistant AI technology developed by Amazon, first used in the Amazon Echo smart speakers developed by Amazon Lab126. It is capable of voice interaction, music playback, making to-do lists, setting alarms, streaming podcasts, playing audiobooks, and providing weather, traffic, sports, and other real-time information, such as news. [3] Alexa can also control several smart devices using itself as a home automation system. Users are able to extend the Alexa capabilities by installing "skills" (additional functionality developed by third-party vendors, in other settings more commonly called apps such as weather programs and audio features).

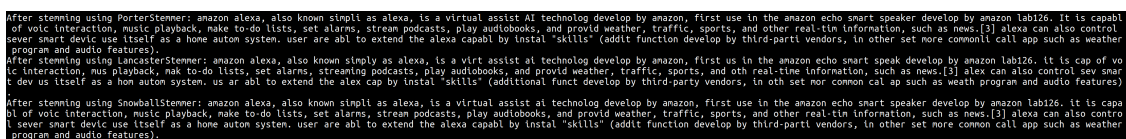
Figure 1.2: Tokenization

1.2.3 Stemming

Stemming refers to reduction of words into their root form. For example, the words do, doing, done all have the same root form 'do'. While processing text containing these words using NLP algorithms, they should be treated the same as they have the same meaning. Thus, it is imperative to reduce all words in a text to their root forms. The NLTK library again offers various stemmers such as:

- Porter Stemmer
- Lancaster Stemmer
- Snowball Stemmer (Supports multiple languages)

Their working is demonstrated below:



After stemming using PorterStemmer: amazon alexa, also known simplly as alexa, is a virtual assist AI technolog develop by amazon, first use in the amazon echo smart speaker develop by amazon lab126. It is capabl of voic interaction, music playback, make to-do lists, set alarms, stream podcasts, play audiobooks, and provid weather, traffic, sports, and oth real-tin information, such as news.[3] alexa can also contro l sever smart devic use itself as a hone autom system. user are abl to extend the alexa capabl by instal "skills" (addit functio develop by third-partl vendors, in oth set more commonl call app such as weath program and audio features).

After stemming using LancasterStemmer: amazon alexa, also known simplly as alexa, is a virt assist ai technolog develop by amazon, first use in the amazon echo smart speak develop by amazon lab126. It is cap of vo ic interaction, mus playback, mak to-do lists, set alarms, stream podcasts, play audiobooks, and provid weather, traffic, sports, and oth real-time information, such as news.[3] alex can also contro l sever smart devic use itself as a hon autom system. us ar abl to extend the alex cap by instal "skills" (addit functio develop by third-partl vendors, in oth set more common cal ap such as weath program and audio features).

After stemming using SnowballStemmer: amazon alexa, also known simplly as alexa, is a virtual assist ai technolog develop by amazon, first use in the amazon echo smart speaker develop by amazon lab126. It is capa bl of voic interaction, music playback, make to-do lists, set alarms, stream podcasts, play audiobooks, and provid weather, traffic, sports, and other real-tin information, such as news.[3] alexa can also contro l sever smart devic use itself as a hone autom system. user are abl to extend the alexa capabl by instal "skills" (addit functio develop by third-partl vendors, in oth set more common call app such as weath program and audio features).

Figure 1.3: Stemming

1.2.4 Lemmatization

Lemmatization, similar to stemming, is the reduction of words into their base form. This base form is achieved by removing morphological differences and utilizing the vocabulary of the language. For example, the words good, better, best could all have the same base form 'good'. The same reasons for utilizing stemming apply to lemmatization, and stemming may not reduce words to forms that lemmatization does, as stemming uses the root form which is always part of the word. NLTK offers the WordNet Lemmatizer, whose working is demonstrated below:

```
After lemmatization using WordNetLemmatizer: Amazon Alexa, also known simply a Alexa, is a virtual assistant AI technology developed by Anazon, first used in the Amazon Echo smart speaker developed by Amazon Lab
126. It is capable of voice interaction, music playback, making to-do lists, setting alarms, streaming podcasts, playing audiobooks, and providing weather, traffic, sports, and other real-time information, such
a news.[3] Alexa can also control several smart device using itself a a home automation system. Users are able to extend the Alexa capability by installing 'skills' (additional functionality developed by third-p
arty vendors, in other setting more commonly called apps such a weather program and audio features).

Differences in words between Stemming and Lemmatization respectively:
bat bat
care caring care
saw saw
feet foot
stripe stripe strip
```

Figure 1.4: Lemmatization

1.3 Vocabulary Construction

The BoW algorithm builds a model using a document-term matrix, constructed by considering the number of occurrences of each word in a document. Using this, each document can be represented as a weighted combination of various words in it. By setting a threshold and choosing words that carry more meaning, a histogram of all the words present in the documents in the form of a feature vector can be made.

Using the Count Vectorizer in the Scikit-learn library, a BoW model is constructed as demonstrated below:

```
There are 78 distinct words in the text. Their occurrences are:
{'amazon': 6, 'alexa': 4, 'also': 5, 'known': 38, 'simply': 37, 'as': 10, 'is': 35, 'virtual': 75, 'assistant': 11, 'ai': 2, 'technology': 65, 'developed': 22, 'by': 15, 'first': 20, 'used': 71, 'in': 31, 'the':
66, 'echo': 25, 'smart': 59, 'speakers': 69, 'labeled': 39, 'it': 36, 'capable': 19, 'of': 45, 'voice': 76, 'interaction': 34, 'music': 42, 'playback': 48, 'making': 41, 'to': 69, 'do': 24, 'lists': 40, 'settin
g': 54, 'alarms': 3, 'streaming': 62, 'podcasts': 50, 'playing': 49, 'audiobooks': 13, 'and': 7, 'providing': 52, 'weather': 77, 'traffic': 70, 'sports': 61, 'other': 46, 'real': 53, 'time': 68, 'information': 3
2, 'such': 63, 'news': 44, 'can': 17, 'control': 21, 'several': 56, 'devices': 23, 'using': 73, 'itself': 37, 'home': 38, 'automation': 14, 'system': 64, 'users': 72, 'are': 9, 'able': 0, 'extend': 26, 'capabili
ties': 18, 'installing': 33, 'skills': 38, 'additional': 1, 'functionality': 29, 'third': 67, 'party': 47, 'vendors': 74, 'settings': 55, 'more': 42, 'commonly': 20, 'called': 16, 'apps': 8, 'programs': 51, 'aud
io': 12, 'features': 27}
```

Figure 1.5: BoW construction

1.4 N-gram modelling

An n-gram is a contiguous sequence of n items from a given sample of text or speech. These items can either be phonemes, syllables, letters, words, or base pairs depending on the application. N-grams are generally collected from a corpus containing text or speech. N-grams are used to generate language models that can predict which word comes next given a history of words.

Bigram, N-gram and Everygram modelling as part of the NLTK library is demonstrated below:

