**About the data:**

The corpus token from Large Movie Review dataset <http://ai.stanford.edu/~amaas/data/sentiment> which contains a set of 25,000 highly polar movie reviews for training, and 25,000 for testing so we take small samples of it to form a corpus of 4712 words.

This sample of one movie review:

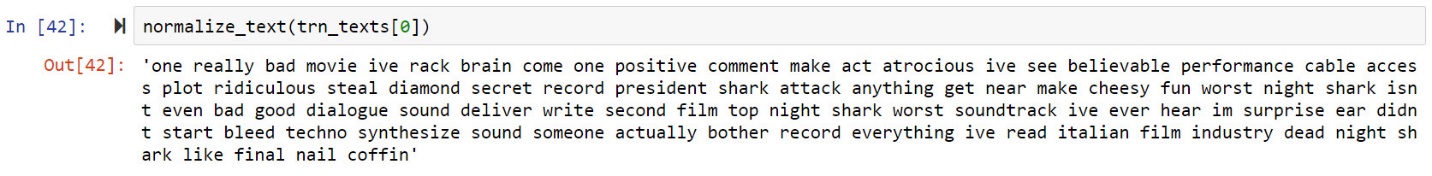
"I grew up on Scooby Doo Where Are You, and I still love it. It is one of my favourite cartoons along with Darkwing Duck, Talespin, Peter Pan and the Pirates and Tom and Jerry. This show though is good for kids, the voices are good(Don Messick and Casey Kasem are perfect as Scooby and Shaggy), the theme tune is tolerable and it has some nice animation. However it is rather disappointing. I normally don't mind Scrappy, but when he appears to be like the main character, it gets annoying fast. Complete with the catchphrase Puppy Power, Scrappy is somewhat more annoying than usual. Also half the gang are missing after the first year, somehow it didn't feel like Scooby Doo. And the jokes and the story lines were in general lame and unoriginal, very little chasing monsters or unmasking the baddies. All in all, not as bad as Shaggy and Scooby Doo:Get a Clue, but this show is disappointing. 4/10 for the animation, voices, theme tune and the fact it is nice for kids. Bethany Cox"

**Text preprocessing:**

we start cleaning the data from stop words, special characters, non\_ascii characters, punctuations and whitespaces, converting all words to lowercase and lemmatizing the words to get small more important words.

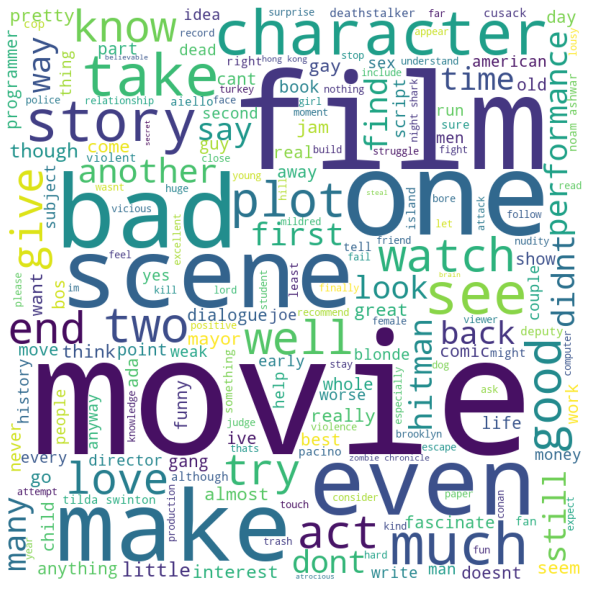
****

This is an example of one text after preprocessing it:



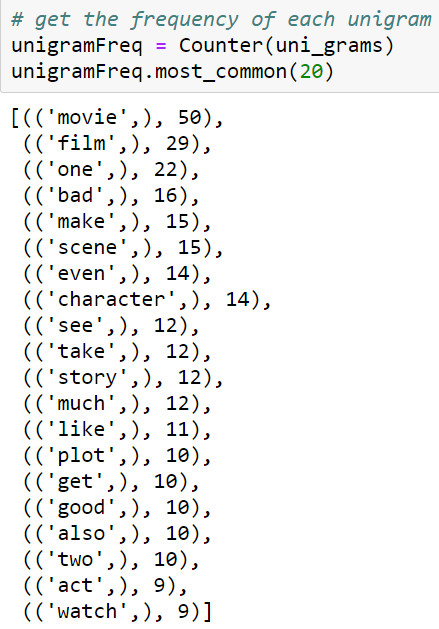
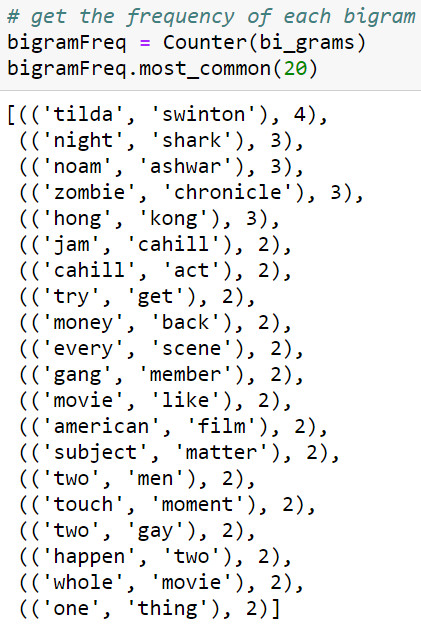
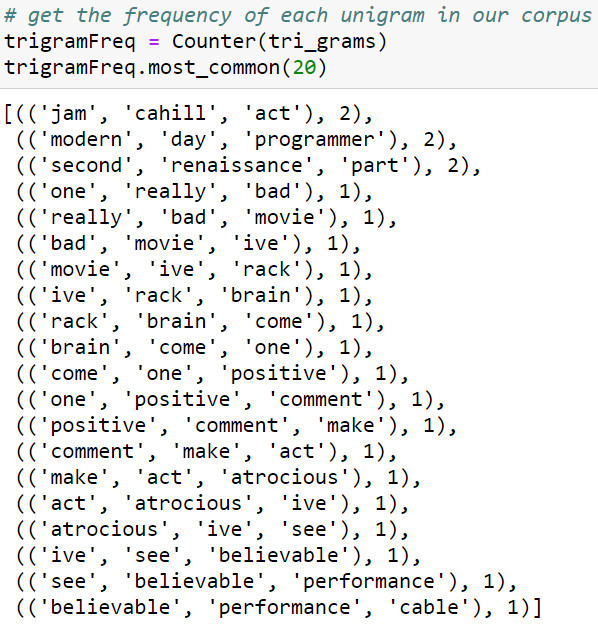
**Visualizing the frequent words:**

Using wordcloud library we visualized the most frequent words in our corpus



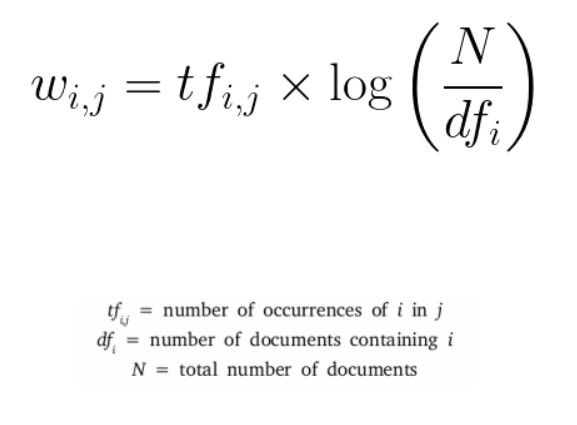
## **Most frequent n\_gram:**

After that we wanted to know what is the top 20 most frequent unigram, bigram and trigrams so we split our data into one words, two words and third words using ngrams function in nltk and counted the words using dictionary collections.

## **TF-IDF**:

Since high frequency words will dominate the vector, specially that it's very sparse, causing quick overfitting. In other words, important low freq words are discarded. This can be treated with TF-IDF using this formula:

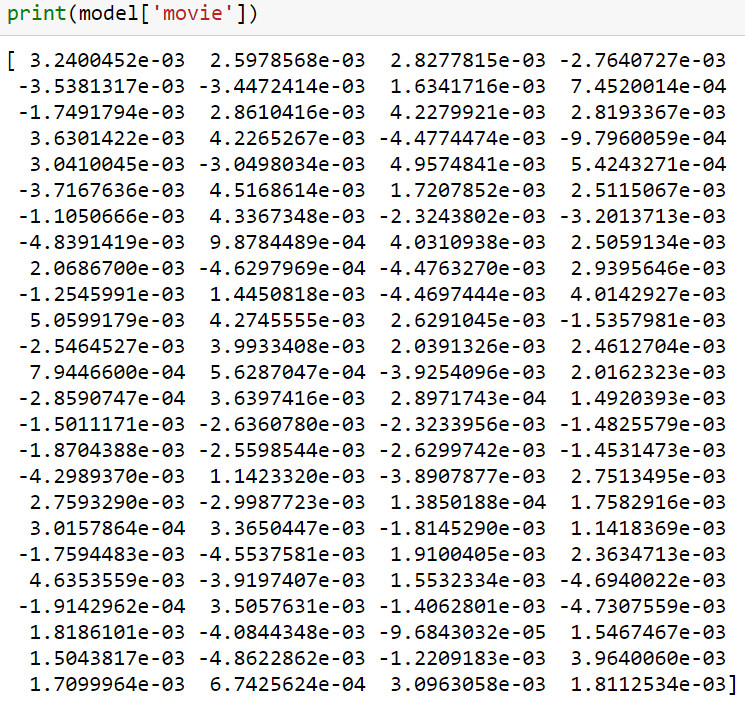


## **Word embedding:**

A word embedding is a class of approaches for representing words and documents using a dense vector representation. It is an improvement over more the traditional bag-of-word model encoding schemes where large sparse vectors were used to represent each word or to score each word within a vector to represent an entire vocabulary. These representations were sparse because the vocabularies were vast and a given word or document would be represented by a large vector comprised mostly of zero values. Instead, in an embedding, words are represented by dense vectors where a vector represents the projection of the word into a continuous vector space. The position of a word within the vector space is learned from text and is based on the words that surround the word when it is used. The position of a word in the learned vector space is referred to as its embedding. Two popular examples of methods of learning word embeddings from text include:

* Word2Vec.
* GloVe.

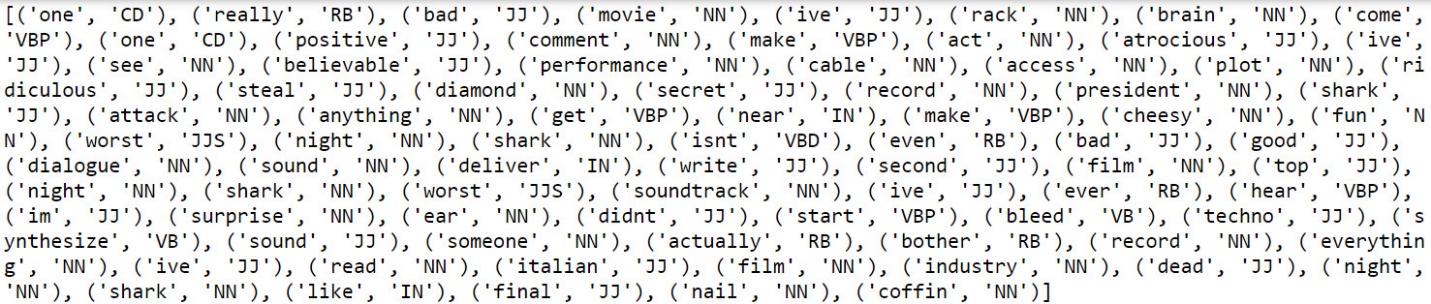
We used CBOW method in Word2Vec class and this is a vector representation for one word in our corpus:



## **PoS tagging:**

Part-of-Speech refers to the different classes a word can belong to: noun, verb, adjective,…etc. The different tags/classes are called tagset, and there's no common standard. They usually encode grammar + tense.

We used NLTK pos tagger forthis task:



**This is a list of some tags and what they mean:**

CC coordinating conjunction

CD cardinal digit

DT determiner

EX existential there (like: “there is” … think of it like “there exists”)

FW foreign word

IN preposition/subordinating conjunction

JJ adjective ‘big’

JJR adjective, comparative ‘bigger’

JJS adjective, superlative ‘biggest’

LS list marker 1)

MD modal could, will

NN noun, singular ‘desk’

NNS noun plural ‘desks’

NNP proper noun, singular ‘Harrison’

NNPS proper noun, plural ‘Americans’

PDT predeterminer ‘all the kids’

POS possessive ending parent‘s

PRP personal pronoun I, he, she

PRP$ possessive pronoun my, his, hers

RB adverb very, silently,

RBR adverb, comparative better

RBS adverb, superlative best