Joel Thompsen
Sebastian Solheim
Gurjot Singh Bains

Assignment report

As humans one of the main ways, we communicate with each other is through the medium of language. I can write a sentence and another person will be able to understand it quickly, assuming we both speak the same language. Machines cannot process text data in raw form like us humans[1]. They need to break down the text information into the format of numbers, so it is readable for the machine. [2] This is where the concept of Bag-of-words comes in.

Making it readable for the machine

Let say you are unsure if you want to watch a new show on Netflix, let's say squid games. Like most people, I will first look up some reviews before committing to watch the show. Here are some examples:

- The show is unrealistic
- The show is interesting and suspenseful
- The show is funny and heart-breaking

From the different reviews, we can not simply feed this data to the machine and ask it to determine if reviews are positive or negative.

Creating Vectors from Text

1. It should not result in a sparse matrix since sparse matrices result in high computation cost[3]
2. We should be able to retain most of the linguistic information present in the sentence [4]

Bag-of-words

BOW model is the simplest form of text representation in numbers. Let's use the examples from above. First, we build a vocabulary from the words above in the reviews. The vocabulary consists of these nine words: "The", "Show", "is", "unrealistic", "interesting", "and" "suspenseful", "funny", "Heart-breaking."

By taking each of the words and marking their occurrence int the show reviews with 1 or 0, we end up with these vectors

Show review 1: (111100000)

---

1 https://techterms.com/definition/machine_language#:~:text=Machine%20language%2C%20or%20machine%20code,they%20only%20recognize%20binary%20data.

2 https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/

3 https://en.wikipedia.org/wiki/Word_embedding#cite_note-1

4 https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/

Joel Thompsen
Sebastian Solheim
Gurjot Singh Bains
Show review 2: (111011100)

Show review 3: (111001011)

Now we can use these vectors and input them into the machine to make it understand.

Drawback of using a Bag-of words model

When adding new words or sentences, it comes at the cost of our vocabulary size, thereby also increasing the length of the vectors. In addition, the vectors would contain many 0s, which would result in a sparse matrix. This is what we are trying to avoid. Lastly the bow model does not factor in any information the ordering of the words nor any grammar information. In conclusion bag of word do have some major issues, especially when we star working on more complex raw text data.

Method:

We took the phrases and given in the assessment and did the various tokenization techniques required in the first few questions, these being tokenization, change to lower case, remove whitespace, remove stop words, remove symbol that is not a number or not an alphabet. We then added all these different methods of tokenisation to their respective lists and printed the list to see how they appear in the terminal. Then after collecting these phrases and the master phrase in the corpus, we ran the tokenisation needed for the Euclidian distance comparison, this being to separate the phrases down to their singular words form and encoding these words using sklearn. Finally, we ran the Euclidian distance comparison in sklearn, comparing our three given phrases to our master statement to determine which is most similar, that being the one with the lowest score.

Results

Results

After running the bag of words model on the different texts, we got the results:27 Unique tokens for Tromsø, 24 Unique tokens for Oslo, 24 Unique tokens for Textmining and 25 Unique tokens for Machine Learning part.

```
Bag of word Tromsø:
{'i': 0, 't': 0, 'a': 0, 'n': 0, 's': 0, 'e': 0, 'r': 0, 'o': 0, 'l': 0, 'c': 0, 'u': 0, 'h': 0, 'd': 0, 'f': 0, 'y': 0, 'g': 0,
'm': 0, 'b': 0, 'w': 0, 'v': 0, '0': 0, 'ˆ': 1, 'ø': 0, 'j': 0, 'p': 0, '1': 0, '2': 0}
We found 27 unique tokens.
```

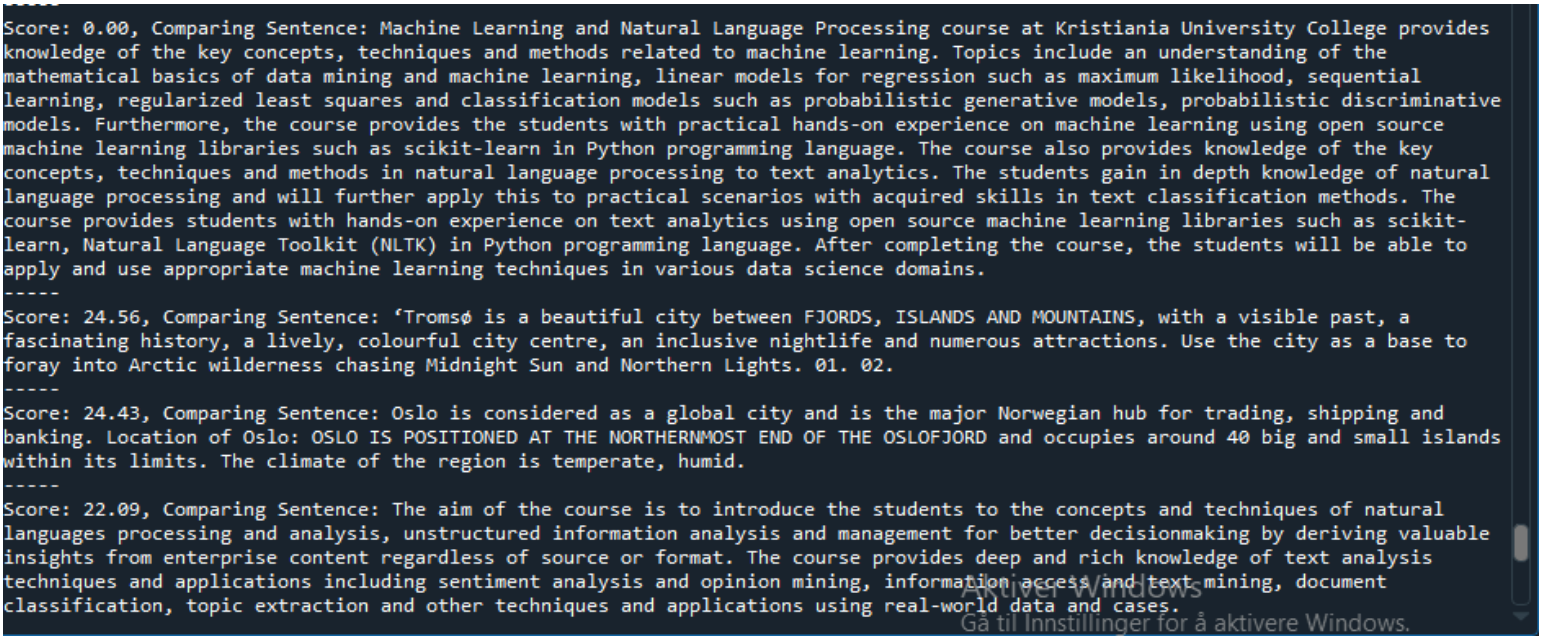
```
Bag of word Oslo:
{'o': 1, 'i': 0, 'n': 0, 's': 0, 'a': 0, 't': 0, 'e': 0, 'd': 0, 'l': 0, 'r': 0, 'h': 0, 'g': 0, 'm': 0, 'c': 0, 'f': 0, 'p': 0,
'b': 0, 'u': 0, 'j': 0, 'w': 0, 'y': 0, 'k': 0, '4': 0, '0': 0}
We found 24 unique tokens.
```

```
Bag of word ML_NLP:
{'e': 0, 'n': 0, 'a': 0, 'i': 0, 's': 0, 't': 0, 'r': 0, 'o': 0, 'l': 0, 'c': 0, 'u': 0, 'h': 0, 'g': 0, 'd': 0, 'p': 0, 'm': 1,
'k': 0, 'f': 0, 'y': 0, 'b': 0, 'v': 0, 'w': 0, 'q': 0, 'x': 0, 'z': 0}
We found 25 unique tokens.
```

```
Bag of word TextMining:
{'n': 0, 'e': 0, 'a': 0, 'i': 0, 't': 1, 's': 0, 'o': 0, 'r': 0, 'c': 0, 'd': 0, 'u': 0, 'l': 0, 'g': 0, 'm': 0, 'h': 0, 'p': 0,
'f': 0, 'y': 0, 'q': 0, 'b': 0, 'v': 0, 'x': 0, 'k': 0, 'w': 0}
We found 24 unique tokens.
```

Joel Thompsen
Sebastian Solheim
Gurjot Singh Bains
Comparing the different sentences.

Scores ended up pretty close to each other, ML ended up with a 0 score because it is the same to the master sentence with zero distance between them.



See Appendix with code for additional results on how the process of cleaning and the original texts.

References:

https://dataaspirant.com/bag-of-words-bow/ (November 1)

https://techterms.com/definition/machine_language#:~:text=Machine%20language%2C%20or%20machine%20code,they%20only%20recognize%20binary%20data. (November 3)

https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/ (November 5)

https://en.wikipedia.org/wiki/Word_embedding#cite_note-1 (November 1)

Lecture material and notes

Videos: (november 1-5)
https://www.youtube.com/watch?v=UFtXy0KRxVI

https://www.youtube.com/watch?v=IKgBLTeQQL8

https://www.youtube.com/watch?v=8Mlc4-3tgzc

https://www.youtube.com/watch?v=aOIHiclLDrc

Joel Thompsen
Sebastian Solheim
Gurjot Singh Bains