CSE556 - Natural Language Processing
Assignment 3 Report

Aditya Chetan (2016217)

September 10, 2018

1 Assumptions

- I have indiscriminately removed punctuation marks without accounting for the semantic meaning of the word. Hence contractions such as, "I've", "She'll" turn to "Ive", "Shell" respectively.
- I have removed all alpha-numeric words completely
- I have also removed all links and emails from the text while preprocessing.
- For the descriminator, I have assumed that the input will be a single sentence
- Also, in the descriminator, I have done `UNK` replacements based on a threshold. So words whose frequencies are below that threshold are replaced with `UNK` and the model is retrained.

2 Methodology

2.1 Language Model Training

The following pipeline is followed:

- First the document is converted to lowercase
- Then it is sentence tokenised using the `nltk` library
- Then from each sentence, all urls and hyperlinks are removes
- After this, from each sentence, the punctuation marks are removed
- Lastly, a dictionary is populated with the appropriate words to keep a record of what words follow what `n-grams`

2.2 Generator

My following pipeline is followed:

- First, depending on the value of `n`, we load the appropriate model.
- Then, the first phrase is selected at random from all the keys in the model that contain `<s>`
- Next, from the values of the selected key in the dictionary, select a random word. Since the words in the value list are non-unique, the probability of getting a more frequent word will be more.
- Next, take the last `n - 1` words of the newly formed sentence and perform the process recursively.
- This goes on until either `</s>` is encountered or a threshold on the number of words in the generated text, which is taken from the user is reached.

2.3 Descriminator

The following pipeline is followed:

- First the input sentence is cleaned.
- If the user wants to replace some less frequent words with `UNK`, the input is accordingly changed and the model is retrained.
- Then feeding it to the descriminator, the counts of each bigram are computed from the model.
- Appropriate Add-k smoothing is done as specified by the user.
- The final probability is calculated by summing up the log likelihoods of each bigram and taking an exponential.
- Finally the results are returned as a `pandas` dataframe.

3 Observations

3.1 Generator

3.1.1 Unigram sentences

Corpus: comp.graphics

```
<s> instructions toaster save surfacemodel also reach from what straightforward
internal microsoft small i as friend the gle replaced as box a of from
holographic hope with workstations id then being of what and the actions
than </s>
```

Corpus: rec.motorcycles

```
<s> large ch they on why to a is order go centerstand quality running
for on and what plus the of has of money there front </s>
```

3.1.2 Bigram sentences

Corpus: comp.graphics

1. <s> but it ought to transfer type of tools </s>

2. <s> noodles from umich in compressed data archives </s>

Corpus: rec.motorcycles

1. <s> interesting content of storage facilities is chain wax out there was a pretty bizzare stuff your ticket california dmv recommended that </s>

2. <s> ama theyre built for your own ed green dod go up for the shaft doesnt so john stafford minnesota state what everybody says for motorcycle enthusiast without a gun and swerving tend to do i will not send to the i take him to what else know </s>

3.1.3 Trigram sentences

Corpus: comp.graphics

1. <s> the vesa local bus video the only person i have the reference point values that come out with slightly different pex implementations </s>

2. <s> section which argues that jpeg can store full color seamlessly tiling photorealistic images for further details </s>

Corpus: rec.motorcycles

1. <s> the bike over to the relay mounted on the security theme is to loud </s>

2. <s> bianchi backstreet suzuki the revolution will not be a bloodsplattered mess and id rather see the plaque that will extend their life considerably </s>