Natural Language Processing

Report for Assignment 1

Vikanksh Nath (MT19AI024)

Visit the above site – choose a corpus you want to work with. *Record why you chose the corpus – who created it and why?*

- We used The NOW (News on the Web) corpus. It contains 9.0 billion words of data from web-based newspapers and magazines from 2010 to the present time. More importantly, the corpus grows by about 140-160 million words of data each month (from about 300,000 new articles), or about 1.8 billion words each year. The Summary of NOW: (Currently) 8.17 billion words | 15 million texts | web pages | 20 different countries | 2010 - last month (Sep 2019) | Growing by 5-6 million words each day.

- **Why NOW -** While other resources like Google Trends show you what people are searching for, the NOW Corpus is the only structured corpus that shows you what is actually happening in the language -- virtually right up to the present time. For example, see the frequency of words since 2010, as well as new words and phrases from the last few years.

Create a corpus of your choice using NLTK – has to have at least 20000 sentences

- **Our Corpus -** We downloaded sample of NOW corpus with more than 2900 news articles. Now on basis of punctuations like full stop, question mark, exclamation mark, we divided the whole plane text into around 77000 sentences. Each sentence has on an average 28 words.

Explain the text processing pipeline adopted by you.

- **Step 1: Sentence Segmentation**

  The first step in the pipeline is to break the text apart into separate sentences (as mentioned already it gives us around 77000 sentences). That gives us these kind of sentences (3 random sentences are shown):

  1. @@ 11241 < p > Sol Yurick , the writer whose 1965 novel " The Warriors " was adapted into a film 14 years later -- which then became one of the best adapted works ever in video gaming -- died this weekend .He was 88 .
  2. < p > Anne : The argument was that while the word hashtag has been around since 2007 , this was the year of the hashtag .
  3. This was the year that hashtag was everywhere in the Twittersphere and beyond @ @ @ @ @ @ @ @ @ @ , making memes go viral .

- **Step 2: First Text Cleaning using regular expression**
  As seen from sentences there are unwanted characters/symbols in our corpus e.g. html tags (< p >), @ *, \t , \n, \r, + etc.

- **Step 3: Word Tokenization**
  We have just splitted apart words whenever there's a space between them. For calculating n-grams we don't need punctuations so for that task we removed it, but for word sense disambiguation task we need punctuations so we created another corpus retaining the punctuations for word sense disambiguation task.

- **Step 4: Predicting Parts of Speech for Each Token (POS)**
  We have used NLTK pos_tag function to find the POS for each token in our corpus. We have considered only four types of POS named adjective, verb, noun and adverb.

- **Step 5: Text Lemmatization**
  We defined a function for Lemmatization to figure out the most basic form or lemma of each word in the sentence. This function takes two arguments tokens and corresponding POS tags. Now using *WordNetLemmatizer* we get the Lemma for each tokens. The output is shown below :-

```
the|None writer|n whose|None novel|None the|None warrior|n be|v adapt|v
into|None a|None film|n year|n later|r which|None then|r become|v one|No
ne of|None the|None best|a adapted|a work|n ever|r in|None video|n gamin
g|n die|v this|None weekend|n he|None be|v yuricks|None work|v itself|No
ne be|v a|None loose|a adaptation|n of|None a|None story|n tell|v year|n
before|None anabasis|n which|None chronicle|v the|None journey|n of|None
greek|a mercenary|n through|None hostile|a territory|n after|None the|No
ne death|n of|None their|None leader|n yuricks|n book|n and|None the|Non
e warrior|n both|None open|a with|None a|None grand|a council|n of|None
street|n gang|n convene|v in|None the|None bronx|n and|None the|None mur
der|n of|None the|None leader|n who|None call|v for|None the|None gather
ing|n cyrus|n a|None direct|a reference|n to|None the|None leader|n of|N
one the|None greek|n in|None anabasis|n but|None the|None story|n then|r
diverge|v significantly|r walter|a hill|v
```

- **Step 6: Identifying Stop Words**
  Some stop words are also removed by manual inspections. After text processing we have only considered the words having more than two characters.

Generate term statistics

- **Vocabulary size with word frequencies**
  Vocabulary size = 55282, we have sorted the vocabulary with frequency, Top High frequency words are shown in table below.

word	say	year	one	make	also	would	time	new	get	people	take
frequency	9604	4364	3903	3433	3362	3209	3166	3119	3021	2917	2667

- **N-grams**

  Total number of bigrams = 634686, Top frequency bigrams shown below in table.

word	per cent	last year	new york	year ago	united state	first time	last week
frequency	541	501	285	282	214	213	207

Total number of trigrams = 812428, Top frequency trigrams shown below in table.

word	india recommend colombia	time india recommend	web time india	around web time	post comment obscene	comment obscene defamatory
frequency	156	155	155	155	88	87

- **POS (Part of Speech) Stats**

For our corpus containing around 77000 sentences and average 28 words per sentence, the total number of repeated words are around 2,156,000. Out of these the POS Counts: {'Noun': 455455, 'Verb': 281830, 'Adjective': 153298, 'Adverb': 75435, 'Others': 566983}

Verify Zipf's law – what is the best fit for your corpus?

- According to Zipf's law f(w) * r(w) = constant.
  f(w) : frequency of word w, r(w) : rank of word w

- Mandelbrot's correction

$$f(w)\ \alpha \frac{1}{[r(w)+\rho]^{\{1+\epsilon\}}}$$

which fitted the language data better. The correction contained two new constants: $0 < \epsilon << 1$ and $\rho >> 1$

We plot the graph between log(frequency) and log(rank)*(1 + ϵ)

For ρ = 10, ϵ = 0.1 the graph is shown in Figure 1.

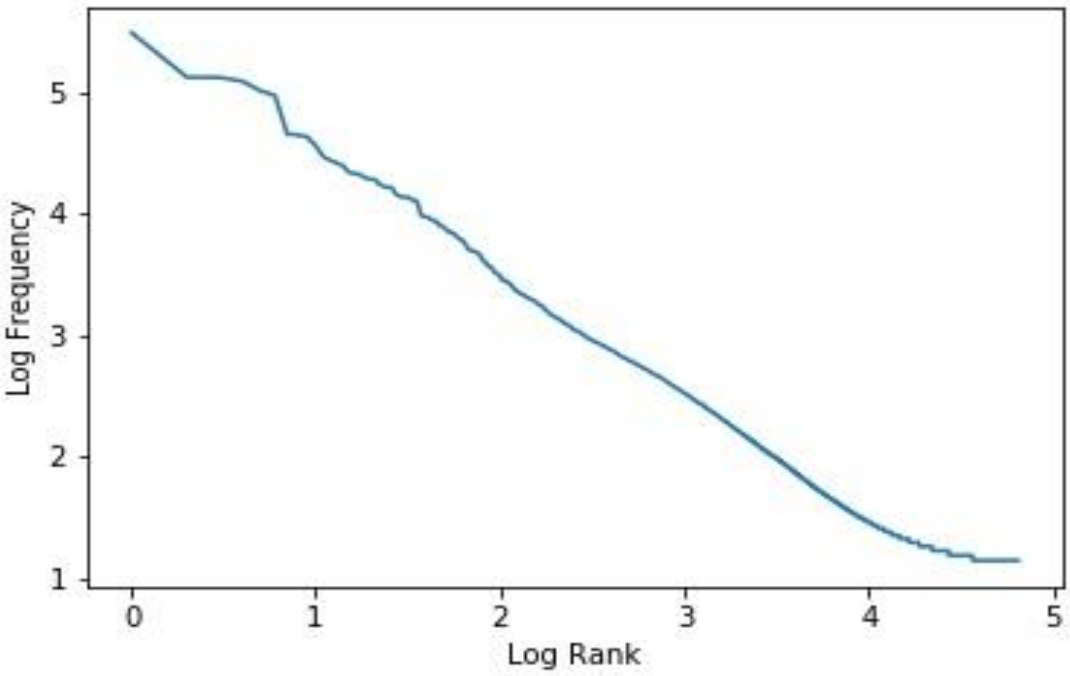

Figure 1: Graph showing word frequency rank distribution

Which set of terms best describe your corpus? How did you arrive at it?

- We used word cloud to identify the best terms in our corpus. 100 significant (key) terms are visualized as shown in Figure 2. Frequency of Noun and adjectives are used to find key terms, as criteria.

Figure 2: Word Cloud of Our Corpus

Using WordNet – find the top 50 most significant yet ambiguous terms in your vocabulary. Explain how you computed significance and why? How did you compute ambiguity?

- The WordNet was created at the University of Princeton for English. It is a semantic network which gives a hierarchical semantic view of the vocabulary of a language. This can be termed as Lexical Ontology too. The nodes in the Wordnet are called the synsets i.e. set of synonyms. Thus, any ambiguous word (The word having more than two meanings) will belong to more than two sysnets.

- Since it is a single text document and after the text pre-processing pipeline, first we simply defined the most frequent words as the most significant ones.

- Out of these significant words we chose those words as most ambiguous which had most number of synsets.

- We selected first top 200 significant words based on frequency and then selected 50 top words having most number of sysnets out of these 200. Thus, we found top 50 ambiguous words for our text.

- The top 55 most ambiguous words for our text are as follows

```
['time', 'first', 'last', 'good', 'way', 'game', 'high', 'part
', 'service', 'life', 'right', 'big', 'home', 'number', 'place
', 'court', 'case', 'work', 'project', 'point', 'report', 'bes
t', 'house', 'man', 'medium', 'second', 'issue', 'young', 'top
', 'change', 'job', 'show', 'book', 'bank', 'end', 'level', 's
mall', 'side', 'name', 'low', 'board', 'bad', 'full', 'record'
, 'order', 'head', 'real', 'press', 'view', 'study', 'free', '
support', 'attack', 'line', 'post']
```

Implement a disambiguation algorithm. Does it serve your purpose? Record your observations with reasons.

- Word sense disambiguation (WSD) algorithm of automatically chooses an appropriate sense for a given word occurrence (target word) in a text (document) out of a set of senses listed in a given dictionary.

- We used the Simplified Lesk Algorithm (SLA) as WSD Algorithm for our task, SLA is a widely used knowledge based WSD algorithm because of its simplicity and speed.

- The steps of SLA are as follows:

```
1 foreach target word W of the document do
2       i = 1
3       while i<N (the context window size) do
4               foreach direction ∈ {left, right} do
5                       w = the word at distance i from W in direction
6                       // Count the overlaps
7                       if w is not a stop word then
8                               foreach sense s of W do
9                                       foreach word u in s (i.e., in definition of the sense s) do
10                                              if u = w then
11                                                      overlap(s) = overlap(s)+1
12              i = i + 1
13      if arg max overlap(s) is unique then
14              Select the sense s for W
15      else
16              Fail to make a decision for W (or use a back-off algorithm to provide it)
```

- **The output of our Lesk Algorithm for ambiguous words "first"," last" is as follows:**

```
                        Word: time
Synset('time.n.05')  ->  We re in a situation now where we ( had nt ) won in a
while , so any time you can score it makes it feel even better . "
 Definition: the continuum of experience in which events pass from the future t
hrough the present to the past

Synset('time.n.03') -> I think it s a great time to have an independent label .
 Definition: an indefinite period (usually marked by specific attributes or act
ivities)

                        Word: last
Synset('end.n.03')  ->  France is the most popular country in the world for tou
rists with almost 85 million visitors last year and the sector employs about tw
o million people .
 Definition: the concluding parts of an event or occurrence
Synset('last.v.01')  ->  It did not surprise me at all when Ramos announced his
own resignation last Monday .
 Definition: persist for a specified period of time
Synset('survive.v.01')  ->  In his regular column last Sunday at The Manila Bul
letin , Ramos wrote : " He ( Duterte ) may claim that to be  more insulting tha
n friendly  to our long established allies is part of his God given  destiny .
 Definition: continue to live through hardship or adversity

Synset('last.n.02')  ->  The accused had robbed an elderly couple last year in
their house at Sector 50 area .
 Definition: the last or lowest in an ordering or series
```

From above results one can see the senses for these ambiguous words have been disambig uated correctly.