QTM 446W

HW2

Gabriel Wang

NLP stands for Natural Language Processing, which is an interdisciplinary study of computer science, artificial intelligence and computational linguistics (Franzosi). It quantitative document analysis process. The basic idea behind NLP is to train machines to recognize patterns of human languages and thus generate statistical results. The better the machine is trained, the more accurate result it will generate. Certain tools always come with NLP. For example: after proper training, NLP can generate a network graph that displays the word co-occurrences. It visualizes what words appears together more often and helps researcher to discover the meaning behind this patter. The tools are all based on ideas like lemmatization and Named-Entity-Recognition.

However, this process may not be very accurate sometimes. There are certain limits. For example, during the process of tokenization and sentence splitting, the meaning of punctuation is always a bummer to NLP. Period is always used at the end of a sentence, but also a lot of other places like decimal number, ellipsis, and abbreviation. There is always some trouble of understanding the true meaning of this period, which may lead to an inaccurate result. Similarly, word sense disambiguation is another challenge which requires the machine to understand the context of the word being used in accurately.

Citation:

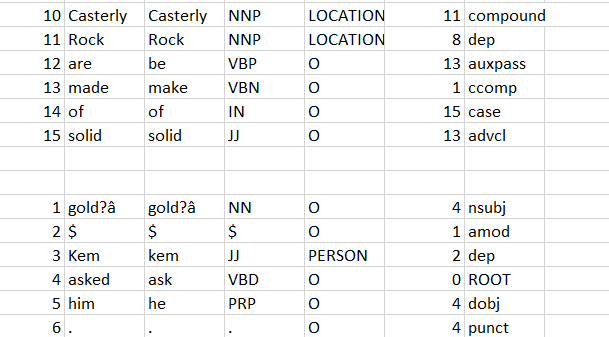
1. Franzosi, Roberto. NLP TIPS files

The CoNLL table result:

I managed to finish part of the Stanford CoreNLP computation of my large corpus, but the process took so long that it literally ran all weekend and didn’t finish. I had to cut the process in the midway and tried to write this assignment based on this partial CoNLL table. The .xlsx PC-ACE transferred from CoNLL file has 1048576 rows in total. I took a screen shot of the last table and realized it barely finished book 3 of 5. 

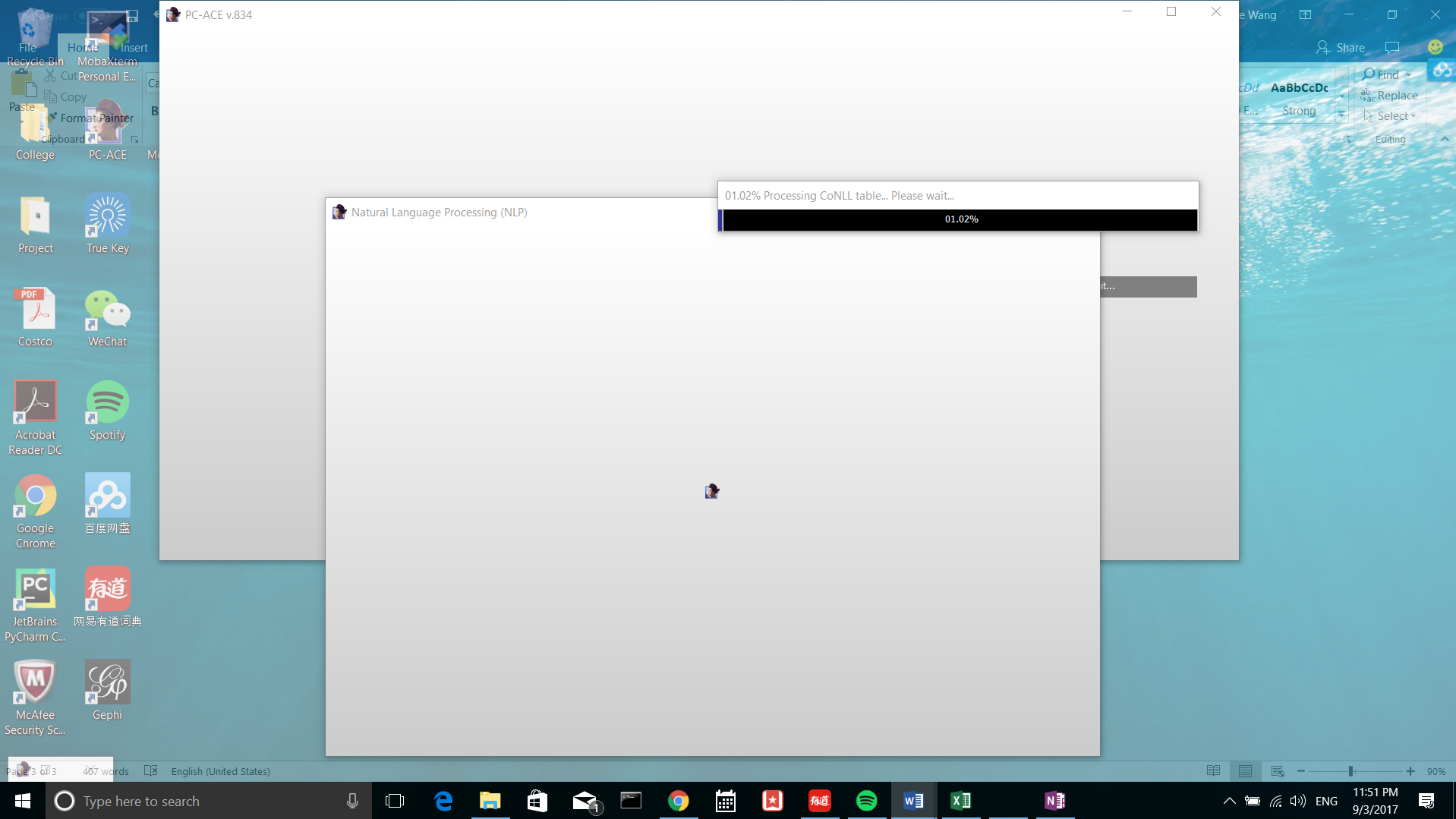
The format of this data is:

1. Id: index of the word in the sentence; eg:1-10
2. Form: the word itself; eg: They, say…
3. Lemma: the lemma or stem; eg: is->be
4. POS: part of speech; eg: VBP verb non-3rd person singular present form;
5. NER: Named-entity recognition; eg: Casterly Rock as LOCATION, Kem as Person;



1. HEAD: index of the syntactic parent; this is part of the semantic parsing (Rodriguez) showing the depth of this word’s in the semantics tree, 0 means the Root;
2. DEPREL: syntactic relationship between HEAD and this word; eg. ROOT means this word is the root;

Improvement: it turns out that the transfer from pdf to txt isn’t accurate enough, that some ‘$’ appear and are recognized as MONEY, this is an improvement I need to work on later.

The Sentence table and KWIX table is still calculating…..