Assignment 1

Name: Serena Yang

I submitted the code with Jupyter Notebook and excel form. Here are only shows the final results for each question.

1. [25 pts] IR Evaluation

Table

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Table

Description automatically generated

Chart, line chart

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

1. [40 pts] Inverted Indexes

Table

Description automatically generated with medium confidence

Text

Description automatically generated with medium confidence

1. i. index AND query

{‘#1’, ‘#2’}

ii. index OR query

{‘#1’, ‘#2’, ‘#3’}

iii. index AND (NOT query)

{}

iv. (search AND query) OR (search AND retrieve)

{‘#1’,’#3’,’#5’}

v. (index OR cluster) AND (engine OR search)

{‘#1’, ‘#5’}

1. [25 pts] Character N-grams to Find Similar Terms

‘informational’ and ‘informally’

A picture containing scatter chart

Description automatically generated

A picture containing text

Description automatically generated

‘informational’ and ‘formalization’

Text

Description automatically generated

A picture containing text

Description automatically generated

‘informational’ and ‘formulation’

A picture containing text

Description automatically generated

A picture containing text

Description automatically generated

1. [10 pts] WordNet

WordNet is a word network. It contains semantic information of words compared with other dictionaries or standard thesauri. Based on the definition of each word, WordNet put the same meaning words into the same group, which is called 'synset.' Besides a word definition, WordNet also provided a brief description for each word using context. When the user searches a word, for example, 'world,' WordNet will return 'world' can be noun and adjective and return the semantic relation word and a context description.

As mentioned above, WordNet is a word-to-word relational dictionary. I think three features of WordNet distinguish it from other types of standard thesauri. Firstly, since the synset is the basic structure for WordNet, users can find more than one appropriate word to present an already known context in the same synset. Assume a WordNet is a dictionary; synset can be the key. Each value can be present as the set of same-meaning words. Secondly, WordNet is not only using synonym relation as its structure concept but also using synonymy, antonymy, hypernymy/hyponymy, meronymy, and entailment word-to-word relation. This structure concept makes the relation between words becomes easier to look up. Third, synset is not equal to a lemma in WordNet, but it is a key because one synset only includes one meaning explanation. However, one lemma could have multiple meanings and statements in a standard dictionary.

WordNet can be used in a few ways in the context of information retrieval. Firstly, it can use for synonyms. When the user has an already known concept, the program can be based on the user description to find the right words. Secondly, in different sentences, the word means different. For example, wordNet can get words' different meanings based on its hyponymy, troponymy, entailment, and meronymy. Also, WordNet can find related words that co-occurrence in a paper to get the word meaning based on the context. Another way is currently using, that combine with context window around six words, the result of using WordNet to eliminate definitions for the polyseme word is optimal. Combining linguistic context and in the similar context of co-occurrence words in WordNet, the accuracy for word sense disambiguation is highest. Semantic relationship in WordNet helps improve search results.