Assignment 1 Marking Scheme

Names: Lingfeng Zhang & Yu Sun

Mark: **94/ 100**

1. Corpus processing: tokenization and word counting [50 points]

a) Submitted a microblog2011_tokenized.txt with the tokenizer's output for the whole corpus. [✓ Y]/N

Included in report the output for the first 20 sentences in the corpus. [✓ Y] / N

b) How many tokens did you find in the corpus? …931,699

How many types (unique tokens) did you have? ………97,008

What is the type/token ratio for the corpus? …0.1041

c) For each token, print the token and its frequency in a file called Tokens.txt Y / [✓ N]

Included the first 100 lines in report. [✓ Y] / N

d) How many tokens appeared only once in the corpus? ……61,310

e) From the list of tokens, extract only words, by excluding punctuation and other symbols.

How many words did you find? ……771,862.

List the top 100 most frequent words in your report, with their frequencies. [✓ Y] / N

What is the type/token ratio when you use only word tokens? ……0.1256

f) List the top 100 most frequent words and their frequencies excluding stopwords. [✓ Y] / N

g) List the top 100 most frequent pairs of two words and their frequencies. [✓ Y] / N

2: Evaluation word embeddings [50 points]

a) Included in the report 8 word embeddings [✓ Y] / N

b) Included the parameters and what mechanism is behind them [✓ Y] / N

c) Table with their results of their evaluations on the 7 benchmark datasets on Similarity [✓ Y] / N

The best scores for Similarity:

	WE	Score
MTurk	ConceptNet	0.7197
MEN	ConceptNet	0.8596
WS353	ConceptNet	0.7546
Rubenstein and Goodenough	ConceptNet	0.9099
Rare Words	ConceptNet	0.5454
SimLex999	ConceptNet	0.6505
TR9856	PDC	0.2073

The best reported average score over all the datasets NA
For what Word Embedding…NA

d) Table with their results of their evaluations on the 4 benchmark datasets on Analogy [✓ Y] / N

The best score for Analogy:

	WE	Score
MSR WordRep	PDC	0.2486
Google_analogy	SG - GoogleNews	0.7476
MSR	PDC	0.7119
SEMEVAL 2012 Task 2	ConceptNet	0.2381

The best reported average score over all the datasets NA
For what Word Embedding…NA

The best reported average score over all the datasets(A,S) 0.5349
For what Word Embedding…ConceptNet