Usama
1606862766

1. Dalam mengerjakan soal ini sama dengan mengerjakan soal yang terdapat pada Tugas 1 lalu. Perbedaannya hanya terdapat pada 1 poin, untuk soal yang sekarang tidak perlu menghitung frekuensi setiap vocab. Oleh karena itu, cukup tandai di *hash* yang kita gunakan jika sudah ada beri nilai *hash* tersebut 1. Hal ini menyebabkan *hash* yang kita miliki bekerja seperti *set*. Setelah selesai memasukkan seluruh *vocab* ke *set*, print seluruh *vocab* yang ada di dalam *hash* ke dalam ***vocabulary_text.txt***.

2. Untuk soal nomor 2 ini, gunakan algoritma yang ada di *paper* "Pande B. P., Dhami H. S. (2011). Application of natural language processing tools in stemming. International Journal of Computer Applications, 27(6): 14–19." untuk lebih jelaskan bisa dilihat di ***T2_1606862766_Usama.p***. Hasil kata yang di*stem* terdapat di ***stemmer_result.txt***. File *txt* tersebut hanya berisikan kata yang distem, jika hasil *stem* sama dengan kata sebelumnya tidak ditampilkan di ***stemmer_result.txt***. Untuk *edit distance* menggunakan **Levenshtein Distance** dan untuk *longest common subsequence*-nya mneggunakan **unigram overlap**.

3.

"lainnya"	Edit Distance	Unigram Overlap	Aturan	
			1	2
lawan	4	3	Y	N
lama	4	3	Y	N
lima	4	3	Y	Y
lain	3	4	Y	Y
Kandidat stem	lima, lain			
Hasil stem	lain			

Terdapat kata yang berhasil diperoleh kata dasarnya seperti kata **lainnya** yang memiliki kata dasar **lain**. Banyak kata yang mengalami mis-*stemming* dikarenakan algoritma yang digunakan untuk melakukan *stemming* tidak cocok

untuk digunakan dalam Bahasa Indonesia. Seperti algoritma **soundex** yang kode fonetik pertamanya merupakan huruf pertama yang dikapitalkan. Sedangkan dalam Bahasa Indonesia terdapat imbuhan di depan kata dasar seperti **dimakan** yang memiliki kata dasar **makan**, namun jika kita menggunakan algoritma yang telah ditentukan kita tidak akan dapat mencari kata dasar dari **dimakan (BXXX)** karena kode fonetiknya sudah pasti berbeda dengan **makan (MXXX)**. Karena dalam melakukan *stemming* bermodalkan kode fonetik yang sama. Mungkin hal ini masih bisa ditolerir jika digunakan untuk bahasa Inggris yang imbuhannya hanya merupakan *suffix* atau bahasa-bahasa lain yang memiliki karakteristik serupa.

4. Berikut merupakan kalimat yang memiliki kode fonetik sama yang ada di **soundex_result.txt** namun banyak memiliki makna yang berbeda.

```
petugas P132
petakan P132
putusan P132
putusannya P132
putus P132
pedesaan P132
patuk P132
pts P132
pudjiastuti P132
potassium P132
pedagangan P132
ptsd P132
patchi P132
pedestrian P132
putuskan P132
padukuhan P132
pdgi P132
patsus P132
pedagang P132
pdts P132
phytosanitary P132
ptsp P132
ppds P132
petisi P132
padjadjaran P132
putusnya P132
```

5. Terdapat di ***edit_distance_unigram_overlap.txt***

Pelajaran yang saya dapatkan saat mengerjakan Tugas 2 ini adalah tidak efektifnya **soundex, edit distance,** dan **longest common subsequences** untuk melakukan *stem*, namun ketiganya akan berpengaruh besar jika digunakan untuk melakukan **correction** penulisan di dalam *query* maupun dokumen.