Nama : Muhamad Achir Suci Ramadhan

NPM : 1706979354

Laporan Tugas 2 Perolehan Informasi

1. Pembuatan Daftar Kata Unik di dalam Tag <TEXT/>

   Daftar kata unik sudah dibuat. Hasilnya dapat dilihat pada file ***vocabulary_list.txt***.

   Asumsi yang digunakan penulis adalah kata-kata di dalam korpus bersifat *case insensitive*, sehingga kata-kata seperti “Apartemen” dan “apartemen” dianggap sama.

2. Penerapan Algoritma *Phonetic Based Stemming Algorithm*

   Algoritma sudah diterapkan. Kodenya dapat dilihat pada file ***T2_1706979354_MuhamadAchirSuciRamadhan.pl***. Hasilnya dapat dilihat pada file ***stemmer_result.txt***.

   Pada bagian penghitungan *edit distance*, karena menggunakan algoritma Levenshtein Distance, maka penukaran dua buah huruf terhitung sebagai dua buah operasi, yaitu satu buah penambahan dan satu buah penghilangan.

3. Analisis Hasil

   Hasil dari algoritma *stemming* pada korpus Bahasa Indonesia yang disediakan menggunakan *Phonetic Based Stemming Algorithm* yang dibuat penulis tidak cukup bagus. Banyak sekali *stem* yang dihasilkan dari suatu kata yang tidak sesuai dengan *stem* yang seharusnya, utamanya untuk kata-kata dengan imbuhan yang merupakan prefiks.

   Hal ini disebabkan pada bagian penentuan kode fonetik menggunakan algoritma *Soundex*, huruf pertama dari suatu kata akan disimpan. Padahal, imbuhan Bahasa Indonesia yang merupakan prefiks sangat sering digunakan, yang berarti seringkali setelah suatu imbuhan ditambahkan, huruf depan suatu kata menjadi berubah. Tetapi, perbedaan huruf depan suatu kata berarti perbedaan kode fonetik dalam algoritma *soundex*. Akibatnya, kata yang diproses dan kata dasarnya tidak masuk ke dalam kelompok yang memiliki kode fonetik yang sama, yang berarti kata yang seharusnya menjadi kata dasar bahkan tidak dijadikan kandidat kata dasar terlebih dahulu oleh algoritmanya.

   Terdapat banyak kata yang mengalami *mis-stemming*, beberapa kata mengalami *understemming*, dan beberapa kata mengalami *understemming*. Namun, tetap ada beberapa kata yang berhasil memperoleh kata dasarnya.

   Berikut adalah contoh pemrosesan suatu kata yang **berhasil** memperoleh kata dasarnya. Kata yang dipilih adalah “ungkapkan” yang memiliki kode fonetik U521. Semua kata yang menjadi kandidat harus memiliki kode fonetik yang sama.

“ungkapkan”	**Edit Distance**	**Unigram Overlap**	**Aturan**	
			I	**II**

ungkapkan	0	9	Y	Y
ungkap	3	6	Y	Y
ungkapan	1	8	Y	Y
ungkapnya	3	7	N	Y
ungkpanya	4	7	N	Y
Kandidat stem	ungkapkan, ungkap, ungkapan			
Hasil stem	ungkap			

Kata-kata seperti “terungkap” dan “mengungkapkan” ada di dalam korpus tetapi tidak masuk ke dalam kandidat karena memiliki kode fonetik yang berbeda, yaitu T652 dan M525.

Berikut adalah contoh pemrosesan kata yang mengalami ***mis-stemming***. Kata yang dipilih adalah “mengungkapkan” dengan kode fonetik M525. Terdapat 179 kata dengan kode fonetik yang sama, tetapi hanya dipilih 5 kata untuk mengilustrasikan prosesnya.

“mengungkapkan”	**Edit Distance**	**Unigram Overlap**	**Aturan**	
			I	**II**
mengungkapkan	0	13	Y	Y
menginap	6	7	Y	Y
menekan	7	6	Y	N
menggemparkan	5	9	N	Y
mengungkap	3	10	Y	Y
Kandidat stem	mengungkapkan, menginap, mengungkap			
Hasil stem	menginap			

Kata “mengungkap” tidak terpilih, dan kata dasar aslinya, yaitu “ungkap”, bahkan tidak masuk ke dalam kandidat karena memiliki kode fonetik yang berbeda, yaitu U521.

Berikut adalah contoh pemrosesan kata yang mengalami ***understemming***. Kata yang dipilih adalah “mewujudkannya” yang memiliki kode fonetik M232. Hanya terdapat 3 kata yang memiliki kode fonetik M232. Semuanya dijadikan kandidat pada ilustrasi berikut.

“mewujudkannya”	**Edit Distance**	**Unigram Overlap**	**Aturan**	
			I	**II**
mewujudkannya	0	13	Y	Y
mewasiatkan	8	6	N	N
mewujudkan	3	10	Y	Y
Kandidat stem	mewujudkannya, mewujudkan			
Hasil stem	mewujudkan			

Kata dasar yang benar adalah “wujud”, namun “wujud” tidak masuk ke dalam kandidat karena memiliki kode fonetik W230.

Berikut adalah contoh pemrosesan kata yang mengalami ***overstemming***. Kata yang dipilih adalah “membawahi” yang memiliki kode fonetik M510. Terdapat 10 kata yang memiliki kode fonetik M510 di dalam korpus yang diberikan.

“membawahi”	**Edit Distance**	**Unigram Overlap**	**Aturan**	
			I	**II**
membiayai	3	7	N	Y
membawahi	0	9	Y	Y
menyuap	6	3	Y	N
membawa	2	7	Y	Y
mampu	7	2	Y	N
menyapa	5	4	Y	N

manibuy	8	3	Y	N
menyapu	6	3	Y	N
mimpi	6	3	Y	N
menepi	6	3	Y	N
Kandidat stem	membawahi, membawa			
Hasil stem	membawa			

Pada kasus ini, *overstemming* yang terjadi dipengaruhi oleh tidak adanya kata "membawah" pada korpus. Selain itu, kata dasar "bawah" memiliki kode fonetik B000 sehingga tidak masuk ke dalam kandidat.

4. Hasil Algoritma Soundex

   Hasil dari algoritma *soundex* dapat dilihat pada file ***soundex_result.txt***.

   Dari algoritma *soundex* yang digunakan, terdapat kata-kata yang memiliki kode fonetik sama, namun sebenarnya memiliki arti yang berbeda, dan sebaliknya, terdapat kata-kata yang sebenarnya memiliki arti atau kata dasar yang sama, tetapi kode fonetiknya berbeda.

   Berikut adalah contoh kata-kata yang memiliki kode fonetik yang sama, namun memiliki arti yang berbeda:

   - M525: menganalogikan, mengingatkan, mengomentari, mengonservasi, dan manusianya.
   - P265: pikiranlah, pasarnya, pacarnya, pusarnya, dan pikirnya.
   - P615: perbincangan, perhubungan, provinsinya, perbandingan, dan perpanjangan.
   - R255: rekomendasi, resminya, rekening, rekaman, dan rekanan.
   - T642: terlokalisasi, terlaksana, terelakkan, trilogi, dan terluka.

   Berikut adalah contoh kata-kata yang memiliki arti atau kata dasar yang sama, tetapi kode fonetiknya berbeda:

   - mengungkap (M525), terungkap (T652), dan ungkap (U521).
   - akhir (A260), berakhir (B626), dan terakhir (T626).
   - ketinggalan (K352), meninggalkan (M552), dan tinggal (T524).

5. Tes Fungsi Edit Distance dan LCS Unigram Overlap yang Dibuat

   Contoh 5 pasangan kata dengan hasil nilai *edit distance* dan *longest common subsequence unigram overlap*-nya dapat dilihat pada file ***edit_distance_unigram_overlap.txt***.