Tugas Besar Struktur Data 2018

Text Similarity (Cek Plagiarisme)

anirahma@jtk.polban.ac.id

Tahapan Pengerjaan

- 31 Mei : Penjelasan spek tugas besar
- 7 Juni : Presentasi dan

  Pengumpulan Deskripsi / Spek/Lingkup tugas yang akan dikerjakan
- 2 – 6 Juli : Demo Program
- 9-11 Juli : Pengumpulan Revisi/perbaikan

- Dekerjakan oleh setiap kelompok dengan jumlah anggota 2-3 orang

Deliverable

- Dokumen Teknis / Laporan, berisi :
  - Deskripsi aplikasi, lingkup, dll
  - Rancangan Kerja dan Pembagian Tugas
  - Rancangan Struktur Data
  - Rancangan Modul
  - Rancangan Interface
  - Rancangan Uji Coba
  - Hasil Uji Coba
  - Lesson Learned
- Source code

Plagiarisme

- KBBI 2008 : Plagiat adalah pengambilan karangan (pendapat, ide dsb) orang lain dan menjadikannya seolah-olah karangan (pendapat) sendiri

- Oxford American Dictionary (2001) : to take and use another person's ideas or writing or inventions as one's own

- Peraturan Menteri Pendidikan RI no 17 tahun 2010

Plagiat adalah perbuatan sengaja atau tidak sengaja dalam memperoleh atau mencoba memperoleh kredit atau nilai untuk suatu karya ilmiah, dengan mengutip sebagian atau seluruh karya dan atau karya ilmiah pihak lain yang diakui sebagai karya ilmiahnya, tanpa menyatakan sumber secara tepat dan memadai

Lingkup Plagiarsime
(Soelistyo, 2011)

1. Mengutip kata-kata atau kalimat orang lain tanpa menggunakan tanda kutip dan tanpa menyebutkan identitas sumbernya.
2. Menggunakan gagasan, pandangan atau teori orang lain tanpa menyebutkan identitas sumbernya.
3. Menggunakan fakta (data, informasi) milik orang lain tanpa menyebutkan identitas sumbernya.
4. Mengakui tulisan orang lain sebagai tulisan sendiri.
5. Melakukan parafrase (mengubah kalimat orang lain ke dalam susunan kalimat sendiri tanpa mengubah idenya) tanpa menyebutkan identitas sumbernya.
6. Menyerahkan suatu karya ilmiah yang dihasilkan dan /atau telah dipublikasikan oleh pihak lain seolah-olah sebagai karya sendiri.

Tipe Plagiarisme (Soelistyo, 2011)

1. Plagiarisme Kata demi Kata (*Word for word Plagiarism*). Penulis menggunakan kata-kata penulis lain (persis) tanpa menyebutkan sumbernya.
2. Plagiarisme atas sumber (*Plagiarism of Source*). Penulis menggunakan gagasan orang lain tanpa memberikan pengakuan yang cukup (tanpa menyebutkan sumbernya secara jelas).
3. Plagiarisme Kepengarangan (*Plagiarism of Authorship*). Penulis mengakui sebagai pengarang karya tulis karya orang lain.
4. *Self Plagiarism.* Termasuk dalam tipe ini adalah penulis mempublikasikan satu artikel pada lebih dari satu redaksi publikasi. Dan mendaur ulang karya tulis/ karya ilmiah. Yang penting dalam *self plagiarism* adalah bahwa ketika mengambil karya sendiri, maka ciptaan karya baru yang dihasilkan harus memiliki perubahan yang berarti. Artinya Karya lama merupakan bagian kecil dari karya baru yang dihasilkan. Sehingga pembaca akan memperoleh hal baru, yang benar-benar penulis tuangkan pada karya tulis yang menggunakan karya lama.

Kategori Plagiat

Sastroasmoro, Sudigdo, 2007)

- Persentase < 30% merupakan plagiat ringan
- Persentase 30%-70% merupakan plagiat sedang
- Persentase > 70% merupakan plagiat berat

Aplikasi Pemeriksaan Plagiat
(Sederhana)

Tahapan Preprocessing

- Case Folding
- Tokenizing
- Stopword Removal
- Stemming

Tekst Preprocessing

Case Folding

- Proses untuk mengubah huruf besar ke huruf kecil dalam dokumen, dilanjutkan dengan menghapus karakter selain ‘a’ sampai ‘z’ (Lestari, et al., 2013).

Tokenizing

- Proses untuk memisahkan kalimat menjadi per kata atau term. Tanda spasi digunakan untuk memisahkan antar term (Lestari, et al., 2013).

Stopword Removal

- *Stopword* adalah kata-kata umum dalam dokumen yang tidak memberikan informasi penting terkait dokumen tersebut (Vijayarani, et al., 2014). Proses ini dilakukan untuk mengurangi jumlah term yang harus diproses.

- Contoh stop word Bahasa Indonesia : yang, di, maka... dst
- Stop word Bahasa Inggris : in, at, so, etc

4 Cara Menghilangkan Stopword

1. The Classic Method

- Metode ini menghilangkan *stopword* yang diperoleh dari daftar atau kamus stopword yang sudah ada.

***2. Methods based on Zipf's Law* (*Z-Methods*)**

- Metode ini menggunakan tiga cara untuk menemukan *stopword* berdasarkan Zipf's Law yaitu : menghilangkan kata yang paling sering muncul (nilai *term frequency* tinggi), menghilangkan kata yang hanya muncul satu kali dan menghilangkan kata dengan nilai *inverse document frequency* yang rendah.

4 Cara Menghilangkan Stopword

***3. The Mutual Information Method* (MI)**

- Metode ini merupakan *supervised method* yang bekerja dengan menghitung *mutual information* antara suatu term dengan kategori dokumen, menghasilkan informasi seberapa penting suatu term terhadap suatu kategori. Apabila suatu term memiliki nilai *mutual information* yang rendah maka term tersebut dihapus karena dianggap tidak mencirikan suatu kategori tertentu.

4. Term Based Random Sampling

- Metode ini mendeteksi sendiri *stopword* yang ada dalam dokumen. Cara kerja metode ini yaitu melakukan iterasi terhadap potongan data yang diambil secara acak, lalu mengurutkan setiap term yang ada berdasarkan nilai *Kullback-Leibler*. *Stopword* dibuat berdasarkan term yang nilai *Kullback-Leibler*-nya rendah.

Stemming

- Stemming adalah proses untuk mengembalikan suatu term ke dalam bentuk dasarnya menggunakan aturan tertentu. Contohnya pada kata bersatu, menyatu, menyatukan, dan kesatuan. Kata-kata tersebut dikembalikan ke bentuk dasarnya yaitu “satu”. Tujuan dari *stemming* yaitu untuk menghilangkan imbuhan, mengurangi jumlah kata, mengurangi waktu proses dan besar memori untuk penyimpanan (Vijayarani, et al., 2014).

Text Similarity

- Setelah preprocessing
- Term frequency
- Cek % similarity

CONTOH

TERM	doc-1	doc-2	doc-3	doc-4	doc-5
term-1	6	7	1	3	
term-2	4	2		4	
term-3			2	5	
term-4	5			3	
term-5	4		3	6	
:					
:					
:					
:					
:					
:					
:					
term-N	2	1		3	

Maka : doc-1 dan doc-4 memiliki similarity yang tinggi