Quiz II Pemrosesan Bahasa Alami
29 November 2018
NIM/Nama :
NIM/Nama :

1. a. Jelaskan pengertian dan perbedaan antara *sparse* vs *dense vector representation*!
   b. Jelaskan pengertian PMI dan beri contoh perhitungan nilai PMI antara beberapa pasangan kata dalam sebuah teks paragraf (bahasa Indonesia)!

   a. *Sparse*: banyak mengandung elemen dengan nilai 0, sementara *dense* kebalikannya (sedikit mengandung elemen dengan nilai 0). Penjelasan lebih lengkap dapat dilihat di buku Speech and Language Processing, chapter 6 vector semantics,
   https://web.stanford.edu/~jurafsky/slp3/6.pdf
   https://web.stanford.edu/~jurafsky/slp3/slides/vector1.pdf
   https://web.stanford.edu/~jurafsky/slp3/slides/vector2.pdf
   b. Definisi PMI terkait dengan kemiripan semantik antar kata: seberapa sering dua kata ditemukan secara bersamaan dalam sebuah dokumen (co-occurred), dibandingkan dengan muncul sendiri-sendiri. Contoh perhitungan PMI antar kata terdapat pada slide berikut: https://web.stanford.edu/~jurafsky/slp3/slides/vector1.pdf , jangan lupa diganti contohnya dengan teks bahasa Indonesia.

2. Beri contoh sebuah aplikasi klasifikasi teks, dan gambarkan (dengan diagram blok) serta jelaskan proses yang terjadi di dalamnya!

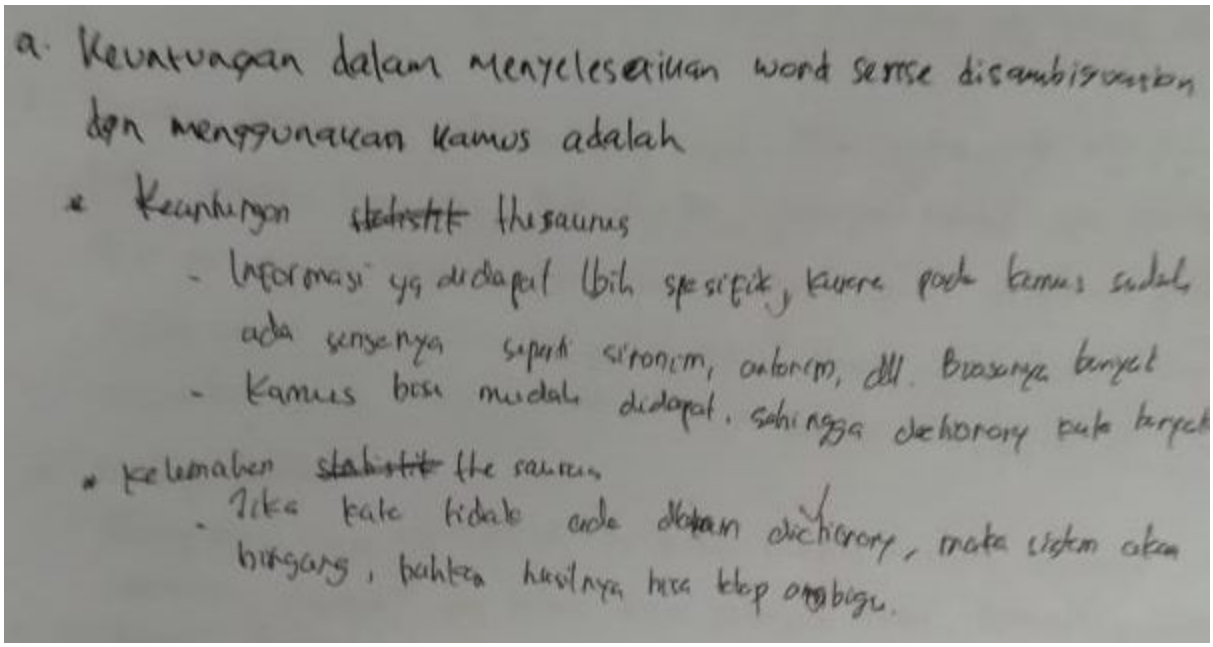
   Contoh (pilih salah 1 saja): spam filtering, klasifikasi sentimen, kategorisasi topik berita, dll

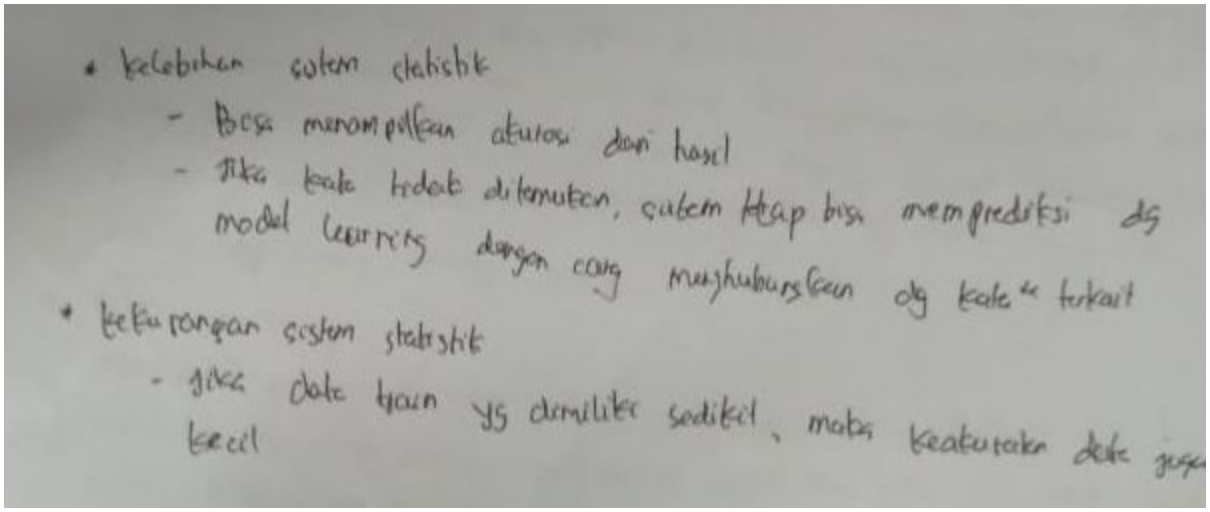
   Diagram blok mengandung komponen:

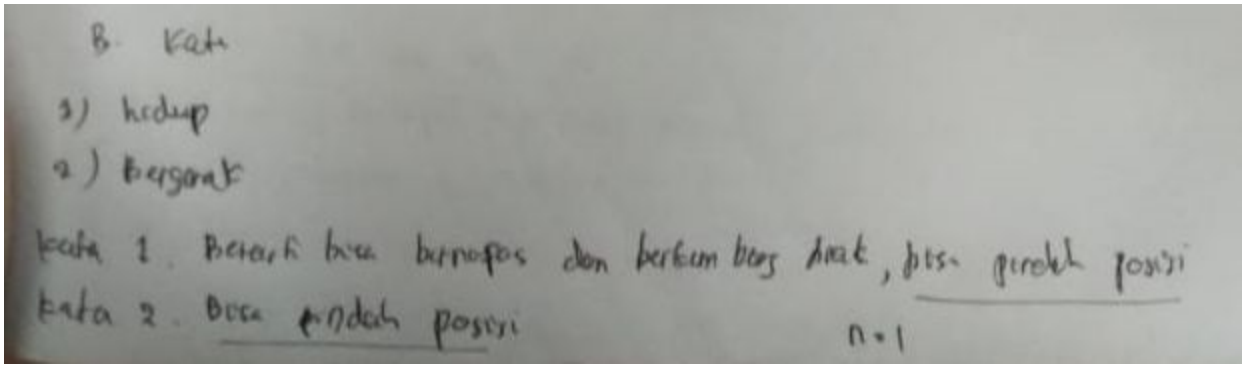
   a. Prapemrosesan data -> berlaku untuk data pembelajaran dan data pengujian, contoh hal yang dilakukan pada saat prapemrosesan adalah: pembersihan data dari karakter-karakter yang dianggap tidak penting, stemming, stop word removal, dan tokenisasi.
   b. Ekstraksi fitur -> berlaku juga untuk data pembelajaran dan data pengujian. Mengambil fitur apa saja yang akan dipakai untuk klasifikasi, misal: vektor count kata per dokumen, vektor tf-idf, banyaknya kata di tiap dokumen, jumlah kata yang termasuk dalam kata kunci sentimen/kelas spesifik lain, dll.
   c. Pembangunan model klasifikasi/pembelajaran -> memasukkan data pembelajaran yang sudah melalui proses prapemrosesan dan ekstraksi fitur ke pengklasifikasi/*classifier*. Keluaran yang diperoleh adalah model klasifikasi.
   d. Pengujian -> klasifikasi data pengujian (yang sudah melalui proses prapemrosesan dan ekstraksi fitur) dengan menggunakan model yang dihasilkan dari data pembelajaran.
   e. Evaluasi, pengukuran kualitas klasifikasi.

Penjelasan di tiap tahap disesuaikan dengan contoh aplikasi yang sudah disebutkan di awal. Jadi misal penjelasan ekstraksi fitur untuk contoh spam filtering akan berbeda dengan penjelasan yang diberikan pada contoh klasifikasi sentimen.

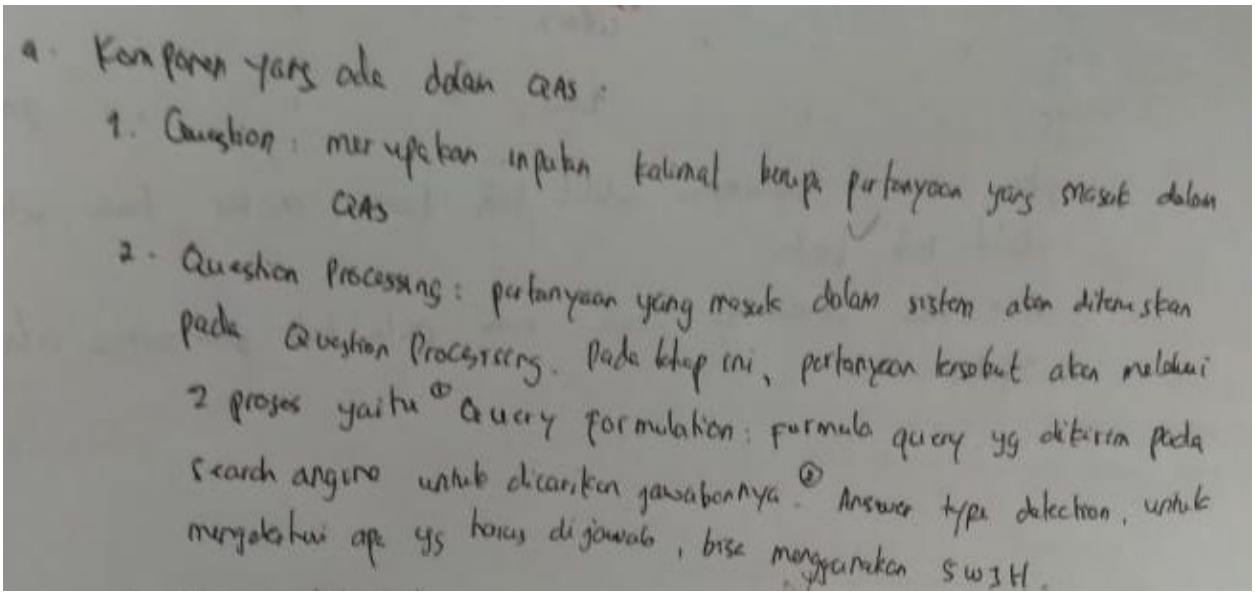
3. a. Jelaskan keuntungan dan kerugian penyelesaian *word sense disambiguation* dengan metode berbasis thesaurus vs metode berbasis statistika!
   b. Berdasarkan informasi definisi *sense* dari KBBI daring (https://kbbi.kemdikbud.go.id), berilah contoh perhitungan kemiripan (*similarity*) antar kata berdasar metode Lesk (dengan penyesuaian bahwa informasi yang tersedia hanya definisi *sense*). Berikan analisis apakah dengan metode perhitungan kemiripan semantik tersebut dapat memberikan hasil cukup bagus (pasangan kata dengan kemiripan semantik tinggi dapat dideteksi berdasar nilai kemiripan yang tinggi, begitu juga sebaliknya).

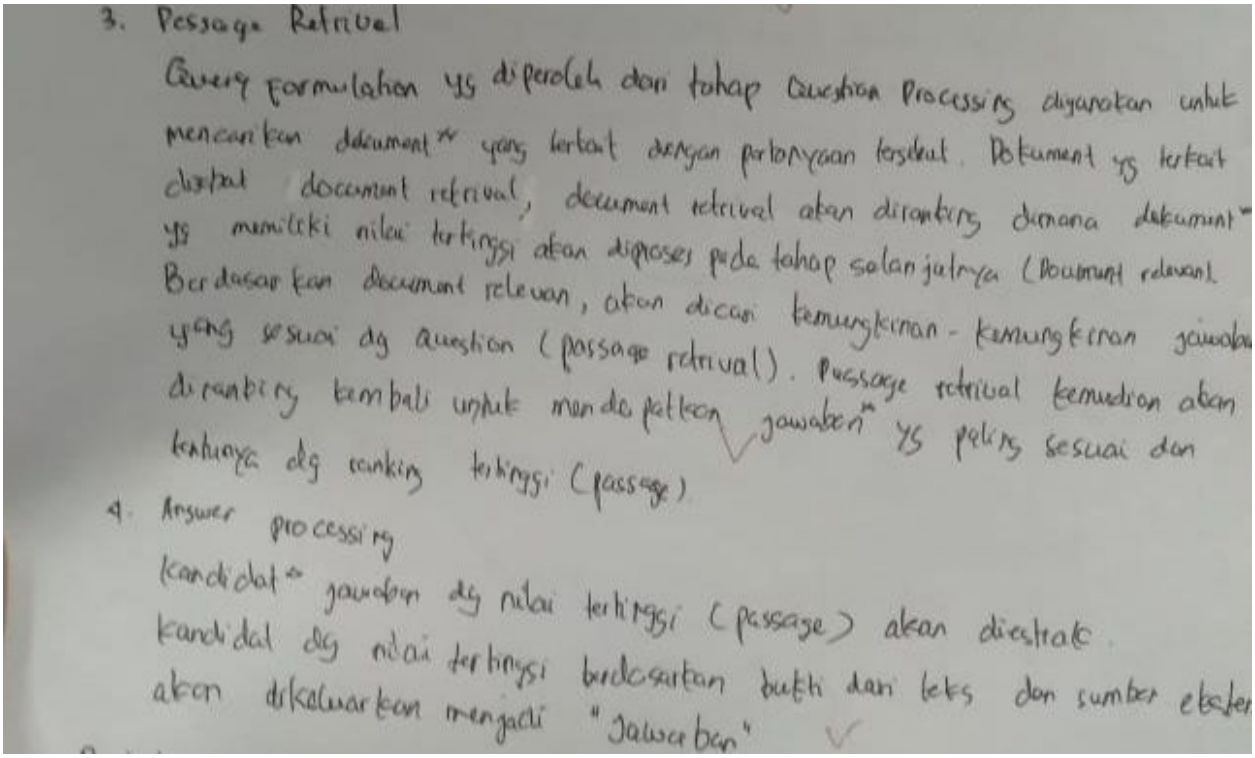

a. Keuntungan dalam menyelesaikan word sense disambiguation dan menggunakan kamus adalah
* Keuntungan ~~statistik~~ thesaurus
- Informasi yg didapat lbih spesifik, karena pada kamus sudah ada sensenya seperti sinonim, antonim, dll. Biasanya banyak
- Kamus bisa mudah didapat, sehingga dictionary cukup banyak
* kelemahan ~~statistik~~ the saurus
- Jika kata tidak ada dalam dictionary, maka sistem akan bingung, hasilnya bisa tdk cukup ambigu.

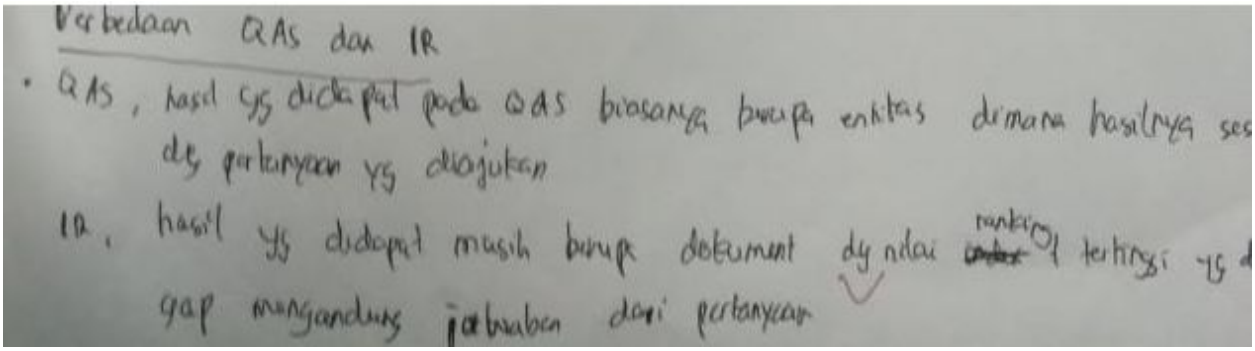

* kelebihan sistem statistik
- Bisa menampilkan akurasi dari hasil
- Jika kata tidak ditemukan, sistem tetap bisa memprediksi dg model learning dengan cara menghubungkan dg kata" terkait
* kekurangan sistem statistik
- jika data train yg dimiliki sedikit, maka keakurasian data juga kecil



4. a. Jelaskan komponen-komponen yang ada dalam sebuah Sistem Tanya Jawab/ Question Answering System berbasis *Information Retrieval*! Apa perbedaan sebuah Sistem Tanya Jawab dan sebuah Sistem *Information Retrieval*?
   b. Beri contoh beberapa aturan/*rule* sederhana yang dapat diterapkan dalam sebuah sistem Ekstraksi Informasi berbasis aturan!







b. Contoh *rule* untuk ekstraksi informasi:

- Pada contoh kasus ekstraksi informasi lowongan pekerjaan, untuk mendapatkan nama perusahaan yang membuka lowongan:
  If rangkaian kata yang diawali dengan token PT. dan mempunyai huruf awal kapital, then Nama Perusahaan