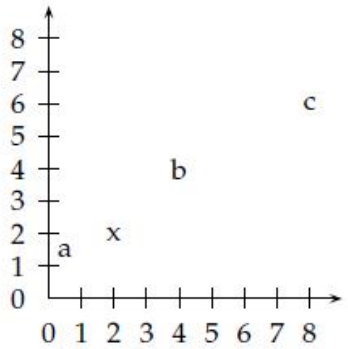
UJIAN AKHIR SEMESTER GANJIL 2011/2012
PENGANTAR TEMU KEMBALI INFORMASI
Jumat, 13 Januari 2011; Pukul 08:15 – 11:00; CATATAN TERTUTUP

Nama Mahasiswa: ..... NIM: .....

---

1. Perhatikan empat dokumen (**a**, **b**, **c**, dan **x**) yang direpresentasikan sebagai vektor seperti gambar di sebelah kanan, dimana **a**'=(0.5 1.5), **b**'=(4 4), **c**'=(8 6), dan **x**'=(2 2).
   a. Jika menggunakan ukuran jarak Euclidean, gambarkan hasil clustering dokumen dengan metode Complete Link. Tunjukkan hasil perhitungan Anda.
   b. Sama dengan pertanyaan (a), tetapi menggunakan ukuran kesamaan Cosine.
   c. Apa yang dapat Anda simpulkan dari hasil pada pertanyaan (a) dan (b)?



2. Perhatikan tiga dokumen berikut:
   (1) *He moved from London, Ontario, to London, England*.
   (2) *He moved from London, England, to London, Ontario*.
   (3) *He moved from England to London, Ontario*.
   Dokumen mana saja yang memiliki representasi fitur yang identik dan yang berbeda jika digunakan model: (a) Multinomial Naive Bayes, dan (b) Bernoulli Naive Bayes. Jika ada perbedaan, jelaskan.

3. Sebutkan dan jelaskan empat jenis *text summarization*.

4. Lakukan pemberian nilai pada passage di bawah ini sehingga dapat diperoleh jawaban yang benar untuk kueri yang berikan. Tunjukkan hasil perhitungan Anda.
   Kueri : *Siapakah Presiden Afganistan yang dipilih pada pemilu 9 Oktober 2004*?
   Passage :

   Terlepas dari segala kekurangannya, Presiden <ORANG>Hamid Karzai</ORANG> yang dipilih melalui pemilu <TANGGAL>Oktober 2004</TANGGAL> mengklaim pemilu parlemen 2005 merupakan tonggak sejarah Afganistan. Setelah 30 tahun berperang, (mengalami) intervensi dan kesengsaraan, hari ini rakyat <ORANG>Afghan</ORANG> bergerak maju. Kami sedang membuat sejarah, tegas <ORANG>Karzai</ORANG> di ibukota <ORGANISASI>Kabul</ORGANISASI>. <LOKASI>Menteri Luar Negeri Afganistan Abdullah</LOKASI><ORANG>Abdullah</ORANG> juga menegaskan bahwa perubahan sedang nerlangsung di negara yang sempat dikuasai rezim Taliban hingga 2001. Perubahan yang dia maksud terutama di bidang pendidikan dan hak perempuan.

5. Untuk melakukan text summarization, digunakan empat fitur f_1, f_2, f_3, dan f_4 untuk setiap kalimat s_i pada suatu dokumen. Dengan menggunakan teknik kombinasi linier, skor setiap kalimat dapat dituliskan sebagai:

$$\text{skor}(s_i) = w_1f_1 + w_2f_2 + w_3f_3 + w_4f_4$$

dimana w_i adalah bobot setiap fitur f_i. Uraikan dengan jelas dan lengkap, bagaimana metode dan teknik untuk mendapatkan nilai bobot setiap fitur tersebut. Asumsikan bahwa Anda diberikan korpus sebanyak 100 dokumen.

6. Salah satu teknik dalam ekspansi kueri adalah menggunakan Thesaurus. Misalkan kata k_1 memiliki sinonim a_1 dan a_2, serta memiliki antonim a_3. Sedangkan kata k_2 memiliki sinonim b_1, dan memiliki antonim b_2. Jika pengguna memberikan kueri: k_1 k_2, maka pilihan kueri apa saja yang ditawarkan kepada pengguna untuk dipilih? Jelaskan.

7. Untuk mengevaluasi algoritme clustering, diujicobakan dua kelompok koleksi dokumen, yaitu A dan B. Kelompok A terdiri atas 8 dokumen, yaitu A1, A2, A3, A4, A5, A6, A7, dan A8. Kelompok B terdiri atas 5 dokumen, yaitu B1, B2, B3, B4, dan B5. Sedangkan kelompok C terdiri dari 4 dokumen, yaitu C1, C2, C3, dan C4. Percobaan pengelompokan seluruh dokumen dilakukan dengan menggunakan dua algoritme, yaitu X dan Y, dan hasilnya adalah:

Algoritme	Kelompok A	Kelompok B	Kelompok C
X	A1, A2, A3, A4, A5, B1	A6, B2, B3, B4, B5, C1	A7, A8, C2, C3, C4
Y	A1, A2, A3, A4, A5, A6, A7, A8	B1, B2, B3, C1, C2, C3	B4, B5, C4

Susunlah *confusion matrix* untuk setiap algoritme, dan hitung akurasinya. Berdasarkan hasil tersebut, algoritme clustering mana yang menurut Anda lebih baik? Jelaskan.

8. Perhatikan paragraf berita berikut:

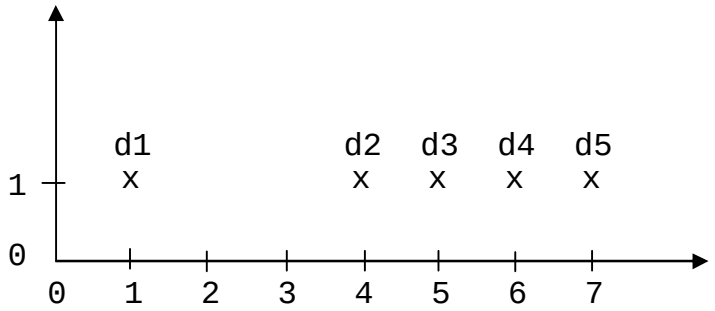
> Hujan di Jakarta tidak deras namun terjadi banjir di beberapa ruas jalan. Akibatnya hanya beberapa kendaraan yang berani melewatinya. Diduga banjir setinggi pinggang orang dewasa itu karena adanya kiriman air dari Bogor. Dugaan ini disampaikan oleh masyarakat yang mengaku terkejut dengan datangnya air yang mendadak itu. Padahal sebelumnya, warga sekitar tidak mendapat peringatan apa pun soal ancaman banjir. Warga curiga, banjir yang menggenangi perumahannya akibat ada pintu air yang tidak dibuka akibat kebakaran di Plumpang. Hujan deras juga diduga sebagai penyebab musibah jebolnya tanggul Situ Gintung.

Diketahui terdapat 100 dokumen dengan frekuensi (***df***: *document frequency*) yang mengandung kata-kata yang diberi garis bawah adalah sebagai berikut:

	air	ancaman	banjir	bogor	hujan	jakarta	kendaraan	tanggul	situ gintung
df	10	5	15	2	5	4	1	2	1

Dengan menggunakan pembobotan tf.idf (gunakan logaritme berbasis 10), ringkaslah paragraf berita tersebut menjadi **hanya dua kalimat**. Tuliskan secara lengkap hasilnya.

9. Perhatikan vektor dari lima dokumen berikut:



Jika lima dokumen tersebut dikelompokkan menjadi dua kelompok dengan menggunakan algoritme cluster single link, bagaimana hasilnya? Dan bagaimana hasilnya jika menggunakan complete link? Mana yang lebih baik? Jelaskan.

10. Uraikan keunggulan dan kelemahan Sphinx-Search dan Lucene. Lakukan pembandingan dan jelaskan mana yang lebih mudah digunakan untuk melakukan penelitian yang menelaah kinerja mesin pencari dengan menggunakan pembobotan TF-IDF, ukuran kesamaan Cosine, dan pencocokan kata berbasis Soundex.