Universitas Kristen Petra Surabaya
Fakultas Teknologi Industri
Program Studi Teknik Informatika

UJIAN TENGAH SEMESTER GENAP 2021 – 2022

Mata Kuliah	:	Natural Language Processing (TF4431)
Hari / Tanggal	:	Kamis / 7 April 2022
Waktu	:	180 menit (10.45 – 13.45 WIB)
Sifat Ujian	:	Open Book
Dosen	:	Dr. Ir. Endang Setyati, M.T.

PETUNJUK: SOAL SESUAI DENGAN CPMK-1

1. (10 point) We are interested in building a language model over a four-word corpus with the words – A, B, C, D, dan E. Consider the following training corpus: ABBCDEEACDEADCCBDEADEADEABBCD. Train a bigram language model on the above, including a <start> and <end> token when needed.
   a. What is P(A|B)?
   b. What is P(B|C)?
   c. What is P(C|D)?
   d. What is P(E|A)?

2. (15 point) What is the most probable next word predicted by the model for the following word sequences? (bigram and trigram)
   a. <s> yesterday . . .
   b. <s> i . . .
   c. <s> i believe . . .
   d. <s> oh I . . .
   e. <s> now I . . .

3. (10 point) Which of the following sentences is better, i.e., gets a higher probability with this model?
   <s> oh I believe </s>
   <s> yesterday </s>
   <s> we she had </s>
   <s> believe in yesterday </s>
   <s> now I need </s>

4. (10 point) Consider again the same training data and the same bigram model. Compute the perplexity of ”<s> oh yesterday”.

5. (15 point) Take again the same training data. This time, use a bigram LM with Laplace smoothing. Give the following bigram probabilities estimated by this model:
   a. P(yesterday|<s>)
   b. P(long|i)
   c. P(i|<s>)
   d. P(far|so)
   e. P(a|is)

   Note : </s> include in vocabulary.

6. (20 point) Let's use Naive Bayes for stanza of a song lyrics identification!
Currently, we have the following corpus of stanza documents that are labeled with their respective stanza-1, 2, 3, 4 of a song lyrics. Unfortunately, the corpus was already converted to all lowercase, which makes this articular task even harder!

Yesterday (paragraph – 1)
All my trouble seemed so far away (paragraph – 1)
Now it looks as though they're here to stay (paragraph – 1)
Oh, I believe in yesterday (paragraph – 1)

Suddenly (paragraph – 2)
I'm not half the man I used to be (paragraph – 2)
There's a shadow hanging over me (paragraph – 2)
Oh, yesterday came suddenly (paragraph – 2)

Why she had to go, I don't know (paragraph – 3)
She wouldn't say (paragraph – 3)
I said something wrong (paragraph – 3)
Now I long for yesterday (paragraph – 3)

Yesterday (paragraph – 4)
Love was such an easy game to play (paragraph – 4)
Now I need a place to hide away (paragraph – 4)
Oh, I believe in yesterday (paragraph – 4)

Why she had to go, I don't know (paragraph – 5)
She wouldn't say (paragraph – 5)
I said something wrong (paragraph – 5)
Now I long for yesterday (paragraph – 5)

Yesterday (paragraph – 6)
Love was such an easy game to play (paragraph – 6)
Now I need a place to hide away (paragraph – 6)
Oh, I believe in yesterday (paragraph – 6)

Using Naive Bayes, determine whether the following sentence is most likely to be classified as paragraf-1 or paragraf-2 or paragraf-3 or paragraf-4 or paragraf-5 or paragraf-6, and calculate the probability that Naive Bayes assigns to that most likely class. For this problem, use Laplace (add-one) smoothing, and ignore any words that do not appear in the corpus.

s = I said something wrong

7. (20 point) Define the hidden Markov model (π, A, B) with the following parameters:
• three states S1, S2, S3, alphabet = {1, 2, 3}.

• $A = \begin{pmatrix} 0 & 0.6 & 0.4 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$

• $\pi = (1, 0, 0)^T$

• $B = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.8 & 0 & 0.2 \\ 0 & 0.4 & 0.6 \end{pmatrix}$

What are all possible state sequences for the following observed sequences O, and what is P(O |(π, A, B))?

~~ SELAMAT MENGERJAKAN, SUKSES UNTUK ANDA ~~

PETUNJUK:

1. Selesaikan semua soal dengan urut. Bila tidak urut dipotong 20 nilai.
2. Kerjakan di word dengan format dan nama file: **UTS-NLP Nama NRP.docx**
3. Subjek: UTS-NLP Nama NRP.
4. Kirim jawaban melalui email ke alamat endang@istts.ac.id.
5. Batas waktu pengerjaan jam 10.45-13.45 WIB. Lebih dari jam tersebut di atas, tidak diperiksa.
6. Bila tidak memenuhi ketentuan dan syarat di atas, berkas tidak diperiksa dan dianggap tidak ikut UTS.