Program Studi Teknik Elektro ITB

Nama Kuliah (Kode) : Praktikum Pemecahan Masalah dengan C (EL2208)
Tahun / Semester : 2019-2020 / Genap
Modul : 9 – Tugas Besar

Naskah Soal Tugas Besar: Tipe 2

Soal tugas besar ini dikerjakan oleh kelompok dengan nomor kelompok genap.

Pembuat Naskah: Muhammad Naufal Thariq

Dalam bidang *computational linguistics* dan probabilitas, sebuah n-gram adalah suatu rangkaian yang terdiri dari n simbol (sesuai dengan namanya) yang didapatkan dari input teks alami (dalam konteks Natural Language Processing). Simbol-simbol ini dapat berupa fonem, silabel, huruf, kata, dan lain-lain. Namun, pada konteks persoalan ini, kita menganggap bahwa n-gram adalah suatu rangkaian yang terdiri dari n kata. Seperti contoh, “saya” adalah 1-gram, “waduh gawat nih” adalah 2-gram, dan “waduh ini 2 minggu tapi kok ga kelar kelar” adalah 9-gram. N-gram dapat digunakan dalam berbagai macam aplikasi. Kebanyakan dari aplikasi tersebut melibatkan Natural Language Processing. Pada persoalan ini, kita akan mencoba memecahkan suatu masalah dalam NLP yang cukup jarang disentuh oleh manusia-manusia di bidang NLP (bahkan beberapa menganggap persoalan ini tidak termasuk ranah NLP). Untuk membaca lebih lanjut tentang n-grams, dapat membuka tautan berikut: **https://en.wikipedia.org/wiki/N-gram**.

Pada persoalan ini, kita akan membuat sebuah string yang terdiri dari kata-kata yang random. Namun, kita akan membuat seolah-olah string random tersebut memiliki gaya penulisan yang cukup mirip dengan gaya penulisan manusia. Sebagai contoh, jika kita memberikan program ini sebuah textfile berisi naskah Hamlet karya Shakespeare, maka program ini akan menciptakan kata-kata yang cukup mirip dengan gaya penulisan Shakespeare pada naskah teater tersebut. Sebagai contoh, jika kita memasukkan naskah Hamlet ke dalam program, program akan memunculkan string sebagai berikut:

```
... thine especial safety,- Which we do tender as we dearly grieve For
that which thou hast done,- must send thee hence With fiery quickness.
Therefore prepare thyself. The bark is ready and the wind at help, Th'
associates tend, and everything is bent For England. Ham. For England?
King. Ay, Hamlet. Ham. Good. King. So is it, if thou knew'st our
purposes. Ham. I see a cherub that sees them. But come, for England!
Farewell, dear mother. King. Thy loving father, Hamlet. Ham. My mother!
Father and mother is man and wife; man and wife is one flesh; and so, my
mother. Come, for England! Exit. King. Follow him at foot; tempt him with
speed aboard. Delay it not; I'll have him hence to-night. Away! for
everything is seal'd and done That else leans on th' affair. Pray you
make haste. Exeunt Rosencrantz and Guildenstern] And, England, if my love
thou hold'st at aught,- As my great power thereof may give thee sense,
Since yet thy cicatrice looks raw and red After the Danish sword, and thy
free awe Pays homage to us,- thou mayst not coldly set Our sovereign
process, which imports at full, By letters congruing to that effect, The
present death of Hamlet. Do it, England; For like the hectic in my blood
he rages, And thou must cure me. Till I know 'tis done, Howe'er my haps,
my joys were ne'er begun. Exit. Scene IV. Near Elsinore. Enter Fortinbras
with his Army over the stage. For. Go, Captain, from me greet the Danish
king. Tell him that by his license Fortinbras Craves the conveyance of a
promis'd march Over his kingdom. You know the rendezvous. if that his
Majesty would aught with us, We shall express our duty in his eye; And
let him know so. ...
```

Yaa… walaupun cukup aneh untuk dibaca dan kadang tidak logis, kita dapat melihat bahwa string kata-kata tersebut cukup mirip dengan gaya penulisan Shakespeare. Kita juga harus menyadari bahwa string tersebut di-generate secara random. Oleh karena itu, Anda akan membuat program ini dengan menggunakan konsep n-grams.

Kita dapat membuat randomisasi seperti yang telah dijelaskan di atas menggunakan n-gram. Ingatlah bahwa n-gram pada dasarnya adalah suatu rangkaian yang terdiri dari n kata. **Kita dapat menggunakan n – 1 kata pertama dalam n-gram untuk memprediksi kata terakhir pada n-gram tersebut.** Sebagai contoh, perhatikan kalimat berikut.

Ships at a distance have every man wish on board.

Pertama, kita dapat memilih n yang akan digunakan. Dalam contoh ini, kita akan mencoba n = 2. Dengan menggunakan nilai n tersebut, kita dapat menyusun n-gram sebagai berikut.

Key	Value
Ships at	a
at a	distance
a distance	have
distance have	every
have every	man
every man	wish
man wish	on
wish on	board.
on board.	Ships
board. Ships	at

Perhatikan bahwa string di sini adalah case-sensitive dan titik koma dan tanda baca lainnya diikutsertakan (suatu kata didefinisikan sebagai kumpulan karakter yang diapit oleh dua spasi). Selain itu, kita menggunakan kata pertama untuk value pada key kata-kata terakhir (sehingga ada word-wrapping di kalimat tersebut). Dengan menggunakan n-gram di atas, kita dapat memilih secara random key pertama yang akan ditampilkan. Misalkan kita mencetak dua kata pertama berikut.

distance have

Untuk menentukan kata selanjutnya, kita menggunakan tabel n-gram di atas sebagai lookup tabel. Key yang digunakan adalah 2 kata terakhir (karena ini adalah 2-gram) yaitu “distance have”. Kita dapatkan kata “every” sehingga string hasil adalah berikut.

distance have every

Selanjutnya, kita menggunakan n-gram untuk menentukan kata selanjutnya menggunakan “have every”, yaitu “man”:

distance have every man

dengan menggunakan n-gram di atas, kita selanjutnya mendapatkan:

distance have every man wish

dan seterusnya.

Jika kita memilih untuk mencetak 20 kata secara acak, string yang dicetak adalah sebagai berikut.

… distance have every man wish on board. Ships at a distance have every man wish on board. Ships at a …

Perhatikan bahwa kita menulis elipsis (…) untuk menandakan bahwa tidak ada awalan atau akhiran dari string ini (karena string ini yaa..random).

Keajaiban penggunaan n-grams dapat dilihat dengan lebih jelas ketika kita menggunakan kalimat berikut sebagai referensi.

to be or not to be just be who you want to be or not okay you want okay

Kita dapat mengkonstruksikan n-gram (jika n = 2) sebagai berikut.

Key	value
to be	{or, just}
be or	not
or not	{to, okay}
not to	be
be just	be
just be	who
be who	you
who you	want
you want	{to, okay}
want to	be
not okay	you
okay you	want
want okay	to
okay to	be

Kita bisa melihat bahwa ada beberapa aspek yang membedakan n-gram ini dengan n-gram sebelumnya. Pertama, beberapa key memiliki lebih dari satu value. Ketika kita menentukan value pada key tersebut, kita dapat merandomisasi pemilihan value tersebut. Sebagai contoh, seperti biasa, kita memilih secar acak key yang akan digunakan sebagai string pertama:

or not

Selanjutnya, kita memilih antara menggunakan “to” atau “okay” untuk kata selanjutnya. Misalkan kita secara acak memilih okay, maka string menjadi:

or not okay

Kita dapat melanjutkan string tersebut sampai 23 kata.

… or not okay you want to be or not to be just be who you want okay to be or not to be …

Kita bisa melihat sifat acak dari kalimat di atas. Namun, kita juga dapat melihat gaya penulisan referensi pada kalimat tersebut.

Untuk membuat string acak lebih mirip dengan referensi, kita menggunakan n yang lebih tinggi. Jika kita mengkonstruksi 3-gram berdasarkan kalimat referensi sebelumnya, kita dapatkan n-gram berikut.

Key	value
to be or	not
be or not	{to, okay}
or not to	be
not to be	just
to be just	be
be just be	who
just be who	you
be who you	want
who you want	to
you want to	be
want to be	or
or not okay	you
not okay you	want
okay you want	okay
you want okay	to
want okay to	be
okay to be	or

Kita dapat generate string acak sebagai berikut.

… who you want to be or not to be just be who you want to be or not okay you want okay to …

Jika dibandingkan dengan kalimat sebelumnya, kita bisa melihat bahwa kalimat acak di atas lebih mirip (dan lebih jelas) dengan referensi dibanding kalimat sebelumnya. Jika belum yakin bahwa kenaikan n akan menyebabkan kalimat acak yang semakin mirip dan jelas, berikut merupakan hasil pembuatan 300 kata random dari file referensi hamlet.txt menggunakan n = 4 dan n = 8 secara berurutan.

Penggunaan 4-gram

```
... gross as earth exhort me. Witness this army of such mass and charge,
Led by a delicate and tender prince, Whose spirit, with divine ambition
puff'd, Makes mouths at the invisible event, Exposing what is mortal and
unsure To all that fortune, death, and danger dare, Even for an eggshell.
Rightly to be great Is not to stir without great argument, But greatly to
find quarrel in a straw When honour's at the stake. How stand I then,
That have a father klll'd, a mother stain'd, Excitements of my reason and
my blood, And let all sleep, while to my shame I see The portraiture of
his. I'll court his favours. But sure the bravery of his grief did put me
Into a tow'ring passion. Hor. Peace! Who comes here? Enter young Osric, a
courtier. Osr. Your lordship speaks most infallibly of him. Ham. The
concernancy, sir? Why do we wrap the gentleman in our more rawer breath
Osr. Sir? Hor [aside to Hamlet] Is't not possible to understand in
another tongue? You will do't, sir, really. Ham. What imports the
nomination of this gentleman Osr. Of Laertes? Hor. [aside] His purse is
```

```
empty already. All's golden words are spent. Ham. Of him, sir. Osr. I know you are not fit. Ham. Not a whit, we defy augury; there's a special providence in the fall of a sparrow. If it be now, 'tis not to come', if it be not now, yet it will come: the readiness is all. Since no man knows aught of what he did. Laer. A Norman was't? King. A Norman. Laer. Upon my life, Lamound. King. The very same. Laer. I know him well. He is the broach indeed And gem of all the nation. King. He made confession of you; And gave you such ...
```

Penggunaan 8-gram

```
... thine especial safety,- Which we do tender as we dearly grieve For that which thou hast done,- must send thee hence With fiery quickness. Therefore prepare thyself. The bark is ready and the wind at help, Th' associates tend, and everything is bent For England. Ham. For England? King. Ay, Hamlet. Ham. Good. King. So is it, if thou knew'st our purposes. Ham. I see a cherub that sees them. But come, for England! Farewell, dear mother. King. Thy loving father, Hamlet. Ham. My mother! Father and mother is man and wife; man and wife is one flesh; and so, my mother. Come, for England! Exit. King. Follow him at foot; tempt him with speed aboard. Delay it not; I'll have him hence to-night. Away! for everything is seal'd and done That else leans on th' affair. Pray you make haste. Exeunt Rosencrantz and Guildenstern] And, England, if my love thou hold'st at aught,- As my great power thereof may give thee sense, Since yet thy cicatrice looks raw and red After the Danish sword, and thy free awe Pays homage to us,- thou mayst not coldly set Our sovereign process, which imports at full, By letters congruing to that effect, The present death of Hamlet. Do it, England; For like the hectic in my blood he rages, And thou must cure me. Till I know 'tis done, Howe'er my haps, my joys were ne'er begun. Exit. Scene IV. Near Elsinore. Enter Fortinbras with his Army over the stage. For. Go, Captain, from me greet the Danish king. Tell him that by his license Fortinbras Craves the conveyance of a promis'd march Over his kingdom. You know the rendezvous. if that his Majesty would aught with us, We shall express our duty in his eye; And let him know so. ...
```

Pada tugas ini, Anda akan membuat membuat program yang memberikan string acak berdasarkan referensi sebuah textfile dengan menggunakan konsep n-gram yang telah dipaparkan di atas. Program ini memiliki beberapa ketentuan sebagai berikut.

1. Program diawali dengan main menu yang berisi judul dan deskripsi program. Format menu ini dibebaskan kepada Anda.
2. Program menerima file eksternal berupa referensi penulisan teks. File eksternal ini dapat berupa tulisan yang berbagai macam, dari novel sampai produk hukum. Beberapa file referensi teks yang dapat digunakan sebagai referensi dalam pengujian program dapat diakses pada link **https://bit.ly/2JuXVtc**.
3. Program harus menerima input n dari user. Perhatikan bahwa n harus bernilai 2 atau lebih. Jika n bernilai 1, program ini sama saja dengan mencetak kata secara random (tanpa ada kemiripan dengan referensi).
4. Program harus melakukan parsing kata dalam file eksternal sehingga menghasilkan kata-kata satuan. Kemudian, program membentuk tabel n-gram sebagaimana telah dijelaskan di atas. Sebuah kata didefinisikan sebagai kumpulan karakter yang diapit oleh dua spasi. Oleh karena itu, ‘England;’, ‘says;”-‘, ‘(sings.)’, ‘butt-end’, dan ‘me.’ dianggap sebagai satu kata (tanda baca diikut sertakan).

5. Program menerima jumlah kata random yang hendak dicetak. Jumlah kata random dapat bernilai 2000 atau lebih.
6. Program mencetak hasil string random yang terdiri dari jumlah yang telah dimasukkan user pada prompt sebelumnya. Perhatikan bahwa harus terdapat elipsis sebelum dan sesudah string itu untuk menandakan bahwa string tersebut bukan awal atau akhir tulisan.

```
... string goes here ...
```

7. Setelah mencetak string, prompt untuk menerima jumlah kata random dimunculkan kembali. Jika user memasukkan jumlah kata lagi, lakukan langkah 6.
8. Jika user tidak mau memasukkan jumlah kata random lagi, kembalikan ke prompt pemilihan file referensi. Jika user tidak mau memasukkan file referensi lagi, program selesai mengeksekusi.
9. Hal-hal yang tidak tercantum di naskah ini dapat diasumsikan. Namun, mohon, sampaikan asumsi tersebut kepada asisten saat asistensi atau waktu yang lain.

Perhatian

Pada varian soal ini, diberikan bonus penilaian untuk masing-masing anggota yang melakukan hal berikut:

- Membuat file referensi text sendiri. Hal ini dapat berupa laporan praktikum Anda, curhatan Anda, dan lain-lain. Semakin menarik, semakin besar bonus nilainya.
- File referensi text diberi nama TextReference_NIM.txt dan dimasukkan ke dalam repository yang sama dengan repository pengerjaan utama.
- Tidak semua anggota kelompok harus melakukan ini, bisa beberapa anggota kelompok saja yang ingin mendapatkan nilai bonus.