Mini-Project #1

Instructor: Moontae Lee
Total Points: 200
Team 1: Yun Jung Huang (UIN: 655772191), Hoang Nha Nguyen (UIN: 671491808),
Anusha (UIN:670194968)

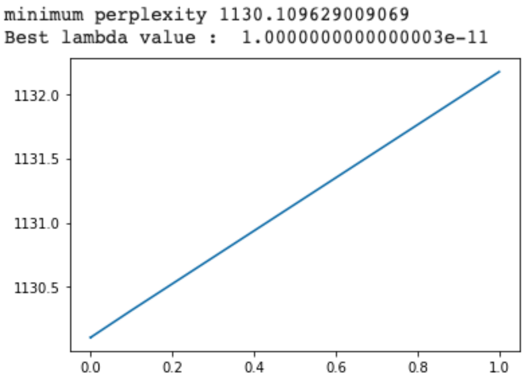
Problem 1: Language Modeling

A. Write a code from the scratch that learns unigram and bigram models on the training data as Python dictionaries. Report the perplexity of your unigram and bigram models on both training data and test data.

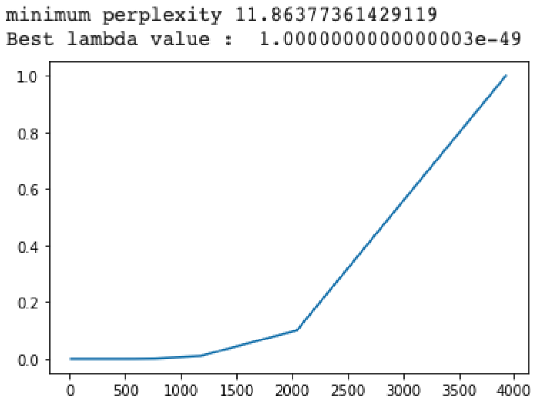
(perplexity)	**Training**	**Testing**
Unigram	1208.3057	inf (infinite for unsmoothing model)
Bigram	64.09040	inf (infinite for unsmoothing model)

.

B. Implement add-λ smoothing method. With varying λ values. Draw a curve that measures your perplexity change over different λ values on the developing data.

<table>
<tr>
<td><b>Unigram Model</b></td>
<td>


```
minimum perplexity 1130.109629009069
Best lambda value :  1.0000000000000003e-11
```

</td>
</tr>
<tr>
<td><b>Bigram Model</b></td>
<td>


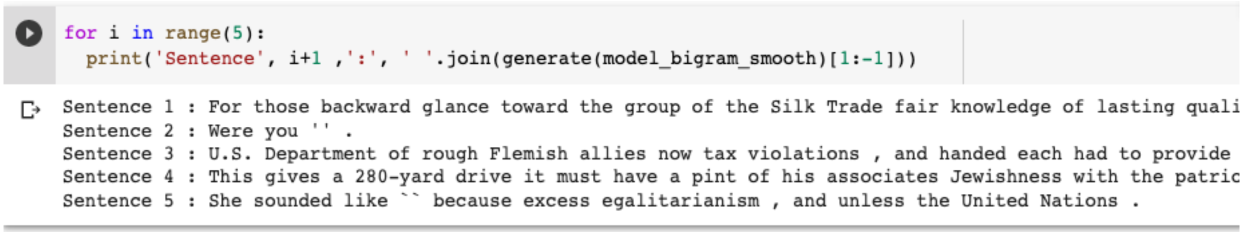
```
minimum perplexity 11.86377361429119
Best lambda value :  1.0000000000000003e-49
```

</td>
</tr>
</table>

C. Pick the best λ value(s) and train again your unigram and bigram models on training data + developing data. Report new perplexity of your unigram and bigram models on the test data.

(perplexity)	Testing
Unigram	1148.5026041516899
Bigram	7.505110653803705

D. Generate random sentences based on the unigram and bigram language models from part (c) Report 5 sentences per model by sampling words from each model continuously until meeting the stop symbol ⟨/s⟩.



```
for i in range(5):
  print('Sentence', i+1 ,':', ' '.join(generate(model_bigram_smooth)[1:-1]))
```

```
Sentence 1 : For those backward glance toward the group of the Silk Trade fair knowledge of lasting quali
Sentence 2 : Were you '' .
Sentence 3 : U.S. Department of rough Flemish allies now tax violations , and handed each had to provide
Sentence 4 : This gives a 280-yard drive it must have a pint of his associates Jewishness with the patric
Sentence 5 : She sounded like `` because excess egalitarianism , and unless the United Nations .
```

Sentence 1 :
For those backward glance toward the group of the Silk Trade fair knowledge of lasting qualities of either -- folded and remote and opening remarks , my ink , enables Fiat 2100 block away than his college girl tried to be accurately for really going to promote school .

Sentence 2 :
Were you '' .

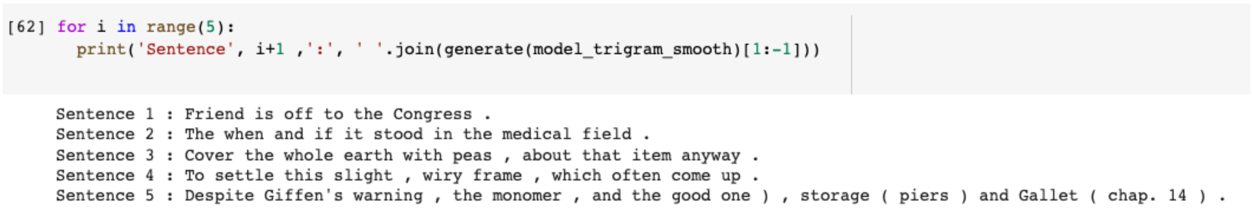
Sentence 3 :
U.S. Department of rough Flemish allies now tax violations , and handed each had to provide each play wasn't no sentry posted at a plane was by jury awarded the Alexander Vasilievitch Suvorov -- political sense , lieutenant general , threatening to budget goes even of ice in small groups) and Mr. Freeman and of complex than we of ultimate purpose .

Sentence 4 :
This gives a 280-yard drive it must have a pint of his associates Jewishness with the patriot in our capabilities '' in great problem.

Sentence 5 :
She sounded like `` because excess egalitarianism , and unless the United Nations .

E. Choose at least one additional extension to implement. The available options are tri- gram, Good-Turing smoothing, interpolation method, and creative handling of unknown words. Verify quantitative improvement by measuring 1) the perplexity on test data; and qualitative improvement by retrying 2) the random sentence generation in part (d).

(perplexity)	Testing
Trigram	1.4553515355342978e-13



```
[62] for i in range(5):
       print('Sentence', i+1 ,':', ' '.join(generate(model_trigram_smooth)[1:-1]))
```

```
Sentence 1 : Friend is off to the Congress .
Sentence 2 : The when and if it stood in the medical field .
Sentence 3 : Cover the whole earth with peas , about that item anyway .
Sentence 4 : To settle this slight , wiry frame , which often come up .
Sentence 5 : Despite Giffen's warning , the monomer , and the good one ) , storage ( piers ) and Gallet ( chap. 14 ) .
```

Sentence 1 : Friend is off to the Congress .

Sentence 2 : The when and if it stood in the medical field .

Sentence 3 : Cover the whole earth with peas , about that item anyway .

Sentence 4 : To settle this slight , wiry frame , which often come up .

Sentence 5 : Despite Giffen's warning , the monomer , and the good one) , storage (piers) and Gallet (chap. 14) .

Problem 2: Application

Pick your open-ended project from one of the following three applications:

- Spell-checking
- Auto-complete

Brainstorm first the interface design. It could be with or without context. Also, it could be in the middle of typing or after finishing the typing. Your approach must be different based on which interface that your group adopts. Relying on your design decision, you could excitingly combine the ideas of character n-grams and word n-grams.

For our application, we decided to create a personalized sentence generator based on the current knowledge we have. (We would love to turn it into a Story generator for the final project.)

The user flow is like this.

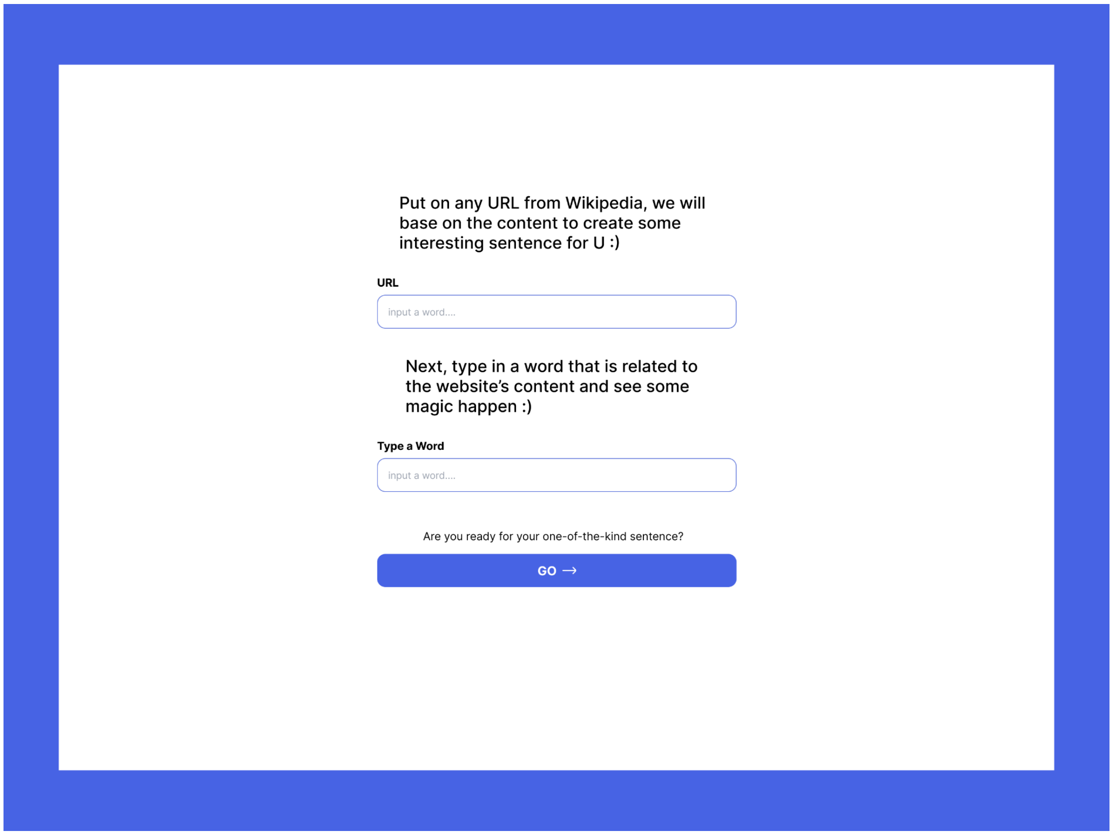
1. Users can paste the webpage of their choice.
2. Users need to type down a word that is related to the webpage's content.
3. Click "GO" CTA button
4. Woa La! The application will return a unique sentence based on the word users typed in step 2

The approach behind the sense is auto-completion using the ngram model.
For this specific application in mini project 1, we use character-based ngram. Since we are using vocabulary-based in problem 1, we would love to give character-based a try in problem 2.

For the web scraping, we used the bs4, BeautifulSoup, package to help us on this. This allows users to have a more fun experience while using the application. They can try out different websites and play around with it!!

Below is our UI brainstorming design. We use Figma to create the interface wireframe.

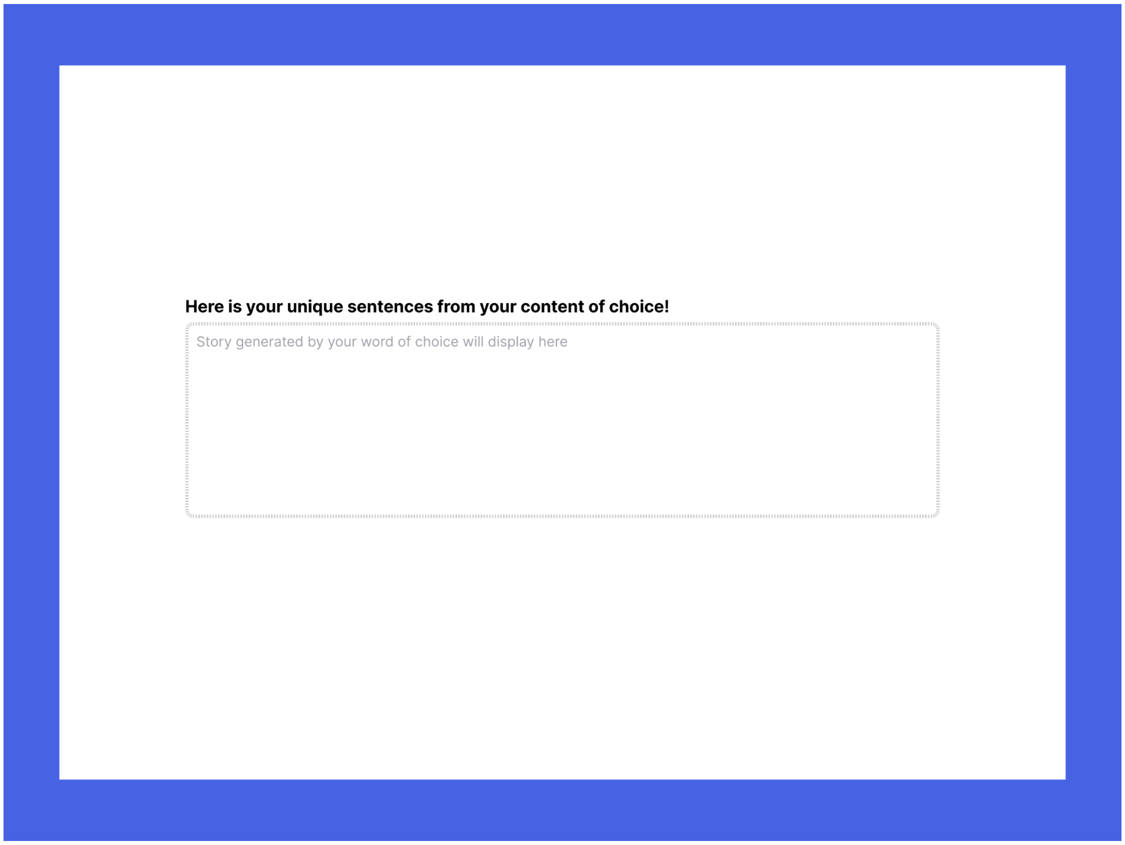
IDS566: Advanced Text Analytics



Put on any URL from Wikipedia, we will base on the content to create some interesting sentence for U :)

Next, type in a word that is related to the website’s content and see some magic happen :)

Are you ready for your one-of-the-kind sentence?



Here is your unique sentences from your content of choice!