UIP HW 2

Monsoon 2019 Ashoka University

Aniruddha Jafa

collaborators: Paul Kurian

1) Corpus:
<s> I am Sam </s>

<s> Sam I am </s>

<s> Sam I like </s>

<s> Sam I do like </s>

<s> do I like Sam </s>

vocabulary: I am Sam like do <s> </s>

a) What is the most probable next word predicted by the model for the following word sequences?

i) <s> Sam …

P(I | Sam) = 3/5

P(</s>|Sam) = 2/5

So the most likely completion is “Sam I”

ii) <s> Sam I do ...

P(like | do) = P(I | do) = 1/2

Thus both completing “Sam I do like” and “Sam I do I” are equally probable **.**

iii) <s>Sam I am Sam ...

P(I | Sam) = 3/5

P(</s>|Sam) = 2/5

So the most likely completion is “Sam I am Sam I”

iv) <s>do I like ...

P(</s> | like) = 2/3

P(Sam | like) = 1/3

So the most likely completion is “do I like </s>”

b) Which of the following sentences is more likely i.e. gets a higher probability with this model?

i)< s>Sam I do I like </s>

P(<s>Sam I do I like </s>)

= e^log(P(Sam | <s>) + P(I | Sam) + P (do | I) + P (I |do) + P(like|I) + P(</s>| like)

[using the fact that logs are additive and *apriori* knowledge that non-of the above probabilities are 0]

= e^log(3/5 ∗ 3/5 ∗ 1/5 ∗ 1/2 ∗ 2/5 ∗ 2/3

= 0.0096

ii)<s>Sam I am </s>

e^log(P(Sam | <s>) + P(I | Sam) + P (am| I) + P(</s> | am))

= e^log(3/5 ∗ 3/5 ∗ 2/5 ∗ 1/2)

[using the fact that logs are additive and *apriori* knowledge that non-of the above probabilities are 0]

= 0.072

iii)<s>I do like Sam I am </s>

e^log(P (I | <s>) + P (do | I) + P(like | do) + P (Sam | like) + P(I | Sam) + P(am | I) + P(<\s> | am)

= e^log (1/5 ∗ 1/5 ∗ 1/2 ∗ 1/3 ∗ 3/5 ∗ 2/5 ∗ 1/2)

= 0.0008

2(a) Attach your sentences grouped by the corpus from which they were generated.

0) ∗ Assumptions ∗

- Important: If n-gram failed to match, we looked at n-1 gram, got the most likely word, and then tried to get back to n grams. e.g. A Tale of Two Cities does not contain "Not two minutes". So we look at 3-grams, get the most likely word *w_i* occurring after the phrase "two minutes", and then try to look at whether or not "two minutes *w_i*" has a match in 4-grams.

- Important : If at a certain point multiple words were equiprobable given a certain sequence (i.e. frequency was the same), a word has been selected randomly. Some of these selections have been described beneath the output tables.

- Length of generated sentence was fixed at 15 for consistency.[1]

- Only lowercase was considered for simplicity. For words that are not proper nouns this has no effect, since capitalisation at the start of sentence needn't be considered since we re already looking at start/end characters. If a word is a proper noun making the first letter lowercase was also not affecting probabilities.

- Commas and special characters were ignored (replaced by whitespace)

- "." was treated as end of sentence replaced with " </s> <s> ". Words like "Mr." and "Mrs." were normalised before this replacement.

[1] In hindsight, this length should have been longer - perhaps at least 20-25.

A)“I suppose”

A.1) Tale of Two cities – “I suppose”

Tale of Two Cities, “I suppose”

1 gram	2 gram	3 gram (not exhaustive)	4 gram (not exhaustive)
<s> i suppose the the the the the the the the the the the the	<s> i suppose </s> <s> i am not a little more than the doctor manette i suppose </s> <s> i am not a little lucie </s> <s> i am	<s> i suppose </s> <s> i am not a word of the night </s> <s> *<s> i suppose </s> <s> i am not a word of it </s> <s> i	<s> i suppose so </s> <s> i am not a traitor </s> <s> at the <s> i suppose so </s> <s> i am not a bad witness </s> <s> the <s> i suppose so </s> <s> i am not a scholar </s> <s> listen to <s> i suppose so </s> <s> i am not a shoemaker by trade </s> <s> <s> i suppose sense in certain quarters you suppose mincing bread and butter may be <s> i suppose sense in certain quarters you suppose mincing bread and butter nonsense </s>

* 2-gram equiprobable completions (not exhaustive – just meant to contextualise some outputs)

- little: more/lucie

* 3-gram equiprobable completions (not exhaustive – just meant to contextualise some outputs)

- word of: the / it

* 4-gram equiprobable completions (not exhaustive – just meant to contextualise some outputs)

- am not a: traitor / scholar / shoemaker / bad
- <s> i suppose: so / sense
- bread and butter: nonsense / and / may

A.2) War and Peace – “I suppose”

War and Peace, “I suppose”

1 gram	2 gram	3 gram (not exhaustive)	4 gram (not exhaustive)
<s> i suppose the the the the the the the the the the the	<s> i suppose that the same time </s> <s> the same time </s> <s> the	<s> i suppose it is not a single word </s> <s> the old prince </s> <s> i suppose it is not a single word of honor </s> <s> the old <s> i suppose it is not a single word of command and the same time <s> i suppose it is not a single word had nicholas alluded to had happened <s> i suppose it is not a single word he said </s> <s> the old <s> i suppose it is not a single word had nicholas been able to find <s> i suppose it is not a single word had nicholas alluded to the emperor <s> i suppose it is not a single word from him </s> <s> the old	<s> i suppose it is one of the most remarkable men of his type see <s> i suppose it is one of the most remarkable men of his squadron that <s> i suppose it is time to harness </s> <s> it was a warm rainy <s> i suppose it is one of the most important the holy of holies of <s> i suppose it is time to harness </s> <s> time is most precious </s>

* 3-gram equiprobable completions (not exhaustive - just meant to contextualise some outputs)

- single word: </s> / of / had / he / from
- word of: honor/command
- alluded to: the/ though / had
- word from : him / me

etc.

* 4-gram equiprobable completions (not exhaustive - just meant to contextualise some outputs)

i suppose it:
completion_possibilities_list [('is', 2), ('has', 1), ('breaks', 1)]

suppose it is:
completion_possibilities_list [('one', 1), ('time', 1)]

the most remarkable:
completion_possibilities_list [('men', 1), ('but', 1)]

men of his:
completion_possibilities_list [('type', 1), ('time', 1), ('squadron', 1), ('company', 1), ('suite', 1)]

harness </s> <s>:
completion_possibilities_list [('it', 1), ('time', 1), ('can', 1)]

etc.

B) “And having got”

B.1) Tale of Two Cities – And having got

Tale of Two Cities, “And having got”

1 gram	2 gram	3 gram (not exhaustive)	4 gram (not exhaustive) - with n-1 grams approach
<s> and having got the the the the the the the the the the the	<s> and having got up and the doctor manette </s> <s> i am not a	<s> and having got his arm </s> <s> i am not a word of it <s> and having got past the two spectators started forward but she sat looking fixedly <s> and having got past the two figures was that </s> <s> i am not <s> and having got past the two brothers looking on at the fountain that although <s> and having got past the two ancient clerks came out of the night </s> <s> and having got past the two brothers crossed the seine denoted the approach of <s> and having got past the two spectators started as he could not be so <s> and having got past the two tall candles he saw the carriage </s> <s> <s> and having got past the two ancient cashiers and shouldered himself into the room <s> and having got past the two tall candles on the road and saw nothing <s> and having got past the two other passengers were close to him </s> <s>	<s> and having got past the two ancient cashiers and shouldered himself into fleet street <s> and having got past the two tall candles on the table and accept your <s> and having got his tools together and all things ready to go down into <s> and having got his tools together and all things on their last roadside with <s> and having got past the two tall candles he saw standing to receive him <s> and having got past the two tall candles on the table </s> <s> i

∗ 3-gram equiprobable completions (not exhaustive – just meant to contextualise some outputs)

- having got: his / past
- the two: spectators / ancient / other / figures / brothers / tall
- two brothers: crossed / looking
- spectators started: forward / as
- word of: it / the
- but she: sat / never / is / so / too / uttered / stayed / was / now / had
- she sat: looking / perfectly / so / down / by / under
- two ancient: clerks / cashiers

etc.

∗ 4 gram equiprobable completions (not exhaustive – just meant to contextualise some outputs)

- past the two: tall / ancient
- two tall candles: on / he
- on the table: and / </s> / in / that / covered / as
- ready to go: down / to / </s>
- having got: past / his

etc.

B.2) War and Peace – “And having got”

War and Peace, “And having got”

1 gram	2 gram	3 gram (not exhaustive)	4 gram (not exhaustive)
<s> and having got the the the the the the the the the the the	<s> and having got up to the same time </s> <s> the same time </ s>	<s> and having got rid of them </s> <s> the old prince </s> <s> the <s> and having got everything ready </s> <s> the old prince </ s> <s> the old <s> and having got rid of her </s> <s> the old prince </s> <s> the <s> and having got warm you daughter dear </s> <s> the old prince </s> <s> <s> and having got warm you daughter of the french army was already in action <s> and having got drunk the day before and had a long time </s> <s> <s> and having got warm then tidied up </s> <s> the old prince </s> <s>	<s> and having got rid of his hat before he ran into prince andrew’s study <s> and having got rid of this young man is the count s son she added <s> and having got rid of his hat before he ran into prince andrew s room <s> and having got rid of this young man was dressed in a threadbare blue <s> and having got rid of this young man was dressed in a purple velvet <s> and having got rid of this young man was spoiled by the depraved ideas

* 3-gram equiprobable completions (not exhaustive – just meant to contextualise some outputs)

- having got: rid / everything / warm / drunk
- rid of: them / her
- french army: was / which
- got warm: then / you / in

etc.

* 4 gram equiprobable completions (not exhaustive – just meant to contextualise some outputs)

- into prince andrew s: study /room
- got rid of: this / his
- young man was: dressed / taking / spoiled / an / </s> / allowed
- dressed in a: threadbare / russian / new / dark purple / black

etc.

C) Not two minutes

C.1) Tale of Two Cities – Not two minutes

Tale of Two Cities, “Not two minutes”

1 gram	2 gram	3 gram (not exhaustive)	4 gram (not exhaustive) - with n-1 grams approach
<s> not two minutes the the the the the the the the the the the	<s> not two minutes </s> <s> i am not a little more than the prisoner	<s> not two minutes and fell asleep on the road </s> <s> i am not <s> not two minutes and as he could not be so much as a honest <s> not two minutes and fell asleep </s> <s> i am not a word of <s> not two minutes and fell asleep on his way to the door </s> <s> <s> not two minutes and as he could not be so </s> <s> i am <s> not two minutes and fell asleep on the road and shaming few good citizens	<s> not two minutes and as there are life and death over the surrounding vulgar <s> not two minutes and as there are life and death in the dominant prison <s> not two minutes and as there are life and death of the city changed <s> not two minutes and as there are life and death of the city </s> <s> not two minutes and as there are life and death of the city arose <s> not two minutes and as there are life and death in the world to

* 3-gram equiprobable completions (not exhaustive - just meant to contextualise some outputs)

- minutes and: fell / as
- fell asleep: on / </s>
- asleep on: the / his
- the road: </s> / and

- much as: a / i / it / thinking / to / suspended / they / get / he

etc.

* 4 gram equiprobable completions (not exhaustive – just meant to contextualise some outputs)

- life and death: over / of / in
- death in the: dominant / world
- of the city: </s> / settling / to / changed / arose / with / gates

etc.

C.2) War and Peace – Not two minutes

War and Peace - “Not two minutes”

1 gram	2 gram	3 gram (not exhaustive)	4 gram (not exhaustive)
<s> not two minutes the the the the the the the the the the	<s> not two minutes later on the same time </s> <s> the same time </s>	<s> not two minutes had elapsed </s> <s> the old prince </s> <s> the old <s> not two minutes had passed the valet who was sitting in the same time <s> not two minutes had passed the chief thing is the matter </s> <s> the <s> not two minutes had passed the examination of the french army was already in <s> not two minutes and natásha were at the same time he had been in <s> not two minutes and natásha were evidently flurried and intimidated by the fact that <s> not two minutes and it was not a single word had nicholas alluded to <s> not two minutes and it was not a single word of honor and favorite <s> not two minutes had elapsed since the ball </s> <s> the old prince </s>	<s> not two minutes had passed before prince vasíli had to go through that we <s> not two minutes had passed before prince vasíli with head erect majestically entered the <s> not two minutes had passed before prince vasíli had to go through that too <s> not two minutes had passed before prince vasíli had to go on a pilgrimage <s> not two minutes had passed before prince vasíli had to go on a tour <s> not two minutes had passed before prince vasíli with that playful manner often employed< <s> not two minutes had passed before prince vasíli had to go on living without having solved the problems that so absorbed him </s> <s>

* 3-gram equiprobable completions (not exhaustive - just meant to contextualise some outputs)

- two minutes
[('had', 1), ('and', 1), ('which', 1)]

- minutes had
[('elapsed', 1), ('passed', 1)]

- passed the

[('examination', 1), ('church', 1), ('commander', 1), ('last', 1), ('company', 1), ('village', 1), ('chief', 1), ('sitting', 1), ('valet', 1), ('fence', 1)]

- had elapsed

[('</s>', 1), ('after', 1), ('since', 1), ('and', 1)]

- as sitting

[('in', 14), ('on', 14)]

- honor and

[('favorite', 1), ('had', 1), ('napoleon', 1), ('the', 1), ('mark', 1), ('glory', 1), ('dishonor', 1), ('hero', 1), ('her', 1)]

etc.

* 4-gram equiprobable completions (not exhaustive - just meant to contextualise some outputs)

- before prince vasíli

[('had', 1), ('with', 1)]

- prince vasíli with

[('head', 1), ('one', 1), ('no', 1), ('that', 1)]

- had to go

[('through', 3), ('on', 3),

- to go on

[('a', 2), ('and', 2), ('living', 2)]

etc.

2(b) Comment on the quality of the generated sentences with increasing complexity of the language model

- A naive 1-gram approach gives meaningless sentences since the same word is repeated.

- 2 grams are a substantial improvement over 1-grams since context is taken into consideration; however there is a tendency to loop because of the limited context e.g. “<s> not two minutes later on the same time </s> <s> the same time </s>”

- 3 and 4 grams give better sentence completions compared to 2 grams since there is more context. However, since there is more context, the corresponding data is also sparse, and a lot of choices become equiprobable. This means more possible sentences can be generated.

With 3 and 4 gram, higher desired_sentence_lengths give more meaningful sentences e.g.

“<s> not two minutes had passed before prince vasíli had to go on living without having solved the problems that so absorbed him </s> <s> (4 gram, sentence length = 25)

“<s> not two minutes had passed before prince vasíli with head erect majestically entered the …” (4 gram, sentence length = 15)

2(c) Comment on the quality of the generated sentences from Corpus 1 and from Corpus 2. Explain any difference that you observe.

- The phrase “I suppose” is slightly more frequent in War and Peace

but this difference is not large enough to make a large impact.

- The phrase “Not two minutes” is in War and Peace but not in Tale of Two Cities. So for Tale of Two Cities, we had to use the n-1 gram approach (else we would get no corresponding output sentence). Even with this modification, for 4-gram, War and Peace gave more meaningful sentences e.g. “<s> not two minutes had passed before prince vasíli with head erect majestically entered the..” (4 gram, sentence length = 15).

- The phrase “And having got” is in in War and Peace, but not in Tale of two cities. This specifically affects 4-grams, since for Tale of Two Cities we had to use the n-1 gram approach. War and Peace gave better sentences on average e.g. “<s> and having got rid of his hat before he

ran into prince andrew’s study” (4 gram, sentence length = 15)

- The conclusion one can draw is that **the corpus matters**, and influences the meaningfulness of completions since it provides a specific context for words occurring in the vocabulary **.** Especially in the 4 gram case, if the input sentence belongs to the corpus, there is a greater likelihood to getting better sentence completions since there is more context to draw from (one doesn’t have to resort to a n-1 grams approach)**.**

- Another factor is **corpus length.** On average, a longer corpus means there is more context, and thus will be able to offer better completions more often, especially for higher grams (e.g. 4 grams). Additionally, one will also have to resort to the n-1 gram approach fewer times (since likely input sentences would probably already be present in the corpus)**.**