Joel Goh
Vincent Rimparsurat

HW#1

1.

Penn Treebank Tokenizer is used to tokenize words according to the Penn Treebank which is a corpus of annotated text data that has been used extensively in NLP. One of the big reasons to use this tokenizer is that almost all of our text is in English, so it contains contractions and punctuation that this tokenizer does well at handling. There are two big differences between the two tokenizers; the first is that the Treebank tokenizer splits words better compared to the GPT2 tokenizer since the GPT2 tokenizer splits words in two more commonly than the NLTK tokenizer; this also means that GPT2 has slightly more tokens for each set. The second difference is that the GPT2 tokenizer contains special characters such as the Ġ to indicate the beginning of a word and Ċ for breaks. The first 200 tokens of the untokenized test set can be seen below.

= Robert <unk> = Robert <unk> is an English film , television and theatre actor . He had a guest @-@ starring role on the television series The Bill in 2000 . This was followed by a starring role in the play Herons written by Simon Stephens , which was performed in 2001 at the Royal Court Theatre . He had a guest role in the television series Judge John <unk> in 2002 . In 2004 <unk> landed a role as " Craig " in the episode " Teddy 's Story " of the television series The Long Firm ; he starred alongside actors Mark Strong and Derek Jacobi . He was cast in the 2005 theatre productions of the Philip Ridley play Mercury Fur , which was performed at the Drum Theatre in Plymouth and the <unk> <unk> Factory in London . He was directed by John <unk> and starred alongside Ben <unk> , Shane <unk> , Harry Kent , Fraser <unk> , Sophie Stanton and Dominic Hall .

The 200 tokens from the GPT and NLTK tokenizers are at the bottom of this document.

2.

The results of applying uni-gram, bi-gram, tri-gram, and 7-gram models on the NLTK tokenized corpus and GPT tokenized corpus are shown below

	NLTK Tokenized	GPT Tokenized
Uni-gram	656.589	706.445
Bi-gram	NaN	NaN
Tri-gram	NaN	NaN
7-gram	NaN	NaN

The only model that produced a perplexity result was the uni-gram model with the rest of the models giving NaN because they had n-grams that resulted in ln(0) which is undefined. One thing that we can determine is that all the words in the test set are in the training set as well otherwise the uni-gram would error out with NaN. It makes sense that as n increases even to 2 the perplexity likely becomes NaN since it becomes significantly less likely that there would be a n-gram in the training set that exactly matches the n-gram from the test set. Thus, it is more likely that there is a ln(0) that results in a NaN.

3.
In an attempt to process text with unknown words or contexts, we first try ignoring n-grams that aren't present in the training set. The results are shown below.

	NLTK Tokenized	GPT Tokenized
Uni-gram	656.589	706.445
Bi-gram	41.428	35.257
Tri-gram	10.390	9.593
7-gram	2.600	2.666

The uni-gram matches the one from Q2 since there were no unknown words and calculates the same perplexity. While we now also have perplexity values for the other models, they are unrealistic measurements. A lower perplexity model typically means that the model can characterize the given test input well, but by excluding the unknown values it only calculates the perplexity of values that are known and does not factor in the large number of unknown n-grams in the test set. Thus, the perplexities are artificially low as the unlikeliness of unseen words is not factored in.

3a.
In another attempt to process text with unknown words or contexts, we try implementing Laplace smoothing. The results are shown below.

	NLTK Tokenized	GPT Tokenized
Uni-gram	658.951	707.652
Bi-gram	830.061	591.575
Tri-gram	4616.201	3186.890
7-gram	29022.051	21641.156

These models use a LaPlace smoothing function for unknown words or context. In this situation, the uni-gram gives a minimally higher perplexity; the difference most likely can be attributed to the smoothing function creating a more uniform distribution thus decreasing the probabilities across all the words, slightly increasing the uncertainty in the model. On the other hand, the other 3 models' perplexity now increases instead of decreasing with each n with the tri-gram and 7-gram models having high perplexity. This high perplexity is due to having extremely low probabilities because of the smoothing function. As the value of n increases linearly, the number of possible n-grams increases exponentially and the individual n-grams themselves become much sparser. Combined with the smoothing, the probability of known n-grams decreases to be closer to the probability of unknown n-grams. As the average n-gram likelihood drops, the perplexity increases.

In another sense, Laplace smoothing reserves too much of the probability space for unknown n-grams and leaves too little for known n-grams. This scales with the number of possible n-grams which itself scales exponentially with the value of n.

4.
The pre-trained GPT2 model gives 23.8789 for the Wikitext-2 test set. We can see that this model performs much better compared to the n-gram models; this can be due to two reasons. The first is that the GPT2 model is trained on a much bigger dataset so it has more information to better understand the test data and have a lower perplexity. The second reason is that the GPT2 model uses a transformer architecture, which has a larger and more flexible context window compared to n-grams which are stuck with a small, fixed context.

5.
The perplexity of the models was measured on a set of examples. The results are shown below.

Example	Uni-gram	Bi-gram	Tri-gram	7-gram	GPT2
1	1483.872	3370.214	16849.754	32164.916	18.371
2	3534.026	5235.325	22053.862	33271.000	96.344
3	2795.093	4526.919	19502.307	33271.000	12.528
4	8675.215	11672.437	25910.114	33271.000	139.320
5	8675.215	11672.436	25910.114	33271.000	36.378
6	2294.834	14015.742	29385.834	33271.000	19.555
7	227.231	54341.064	33271.000	NaN	151.078
8	4181.877	8751.601	21834.720	33271.000	18.244
9	4823.907	14843.932	28595.467	33271.000	137.231

10	1139.578	2265.001	14011.544	33271.000	20.022
11	2603.836	4273.913	18647.556	33271.000	14.564

Looking at the table, the GPT2 model performs consistently better than the n-gram models; the smoothed n-gram models perform as expected where the uni-gram has the lowest perplexity with the perplexity increasing with n. Some interesting points are examples 4 and 5, which when tokenized have the same perplexity since the extra spaces are removed, but the spaces greatly increase the perplexity for the GPT2 model. Also, example 7 only has 3 tokens so can not run the 7-gram model. Lastly, all the models at 7-gram give perplexity around 33271 (3-gram for example 7), which seems to be the perplexity that the smoothed models saturate at. If all n-grams in the example are unknown, then the probability for each n-gram is 1/33271 because it becomes 1/|V| since the two count terms become 0. Then, since all the probabilities are the same, the perplexity becomes the exp(-ln(1/|V|) which equals to |V|.

NLTK Tokenizer

=
Robert
<
unk
>
=
Robert
<
unk
>
is
an
English
film
,
television
and
theatre
actor
.
He
had
a
guest
@
-
@
starring
role
on
the
television
series
The
Bill
in
2000
.
This
was
followed
by
a
starring
role
in
the
play
Herons
written
by
Simon
Stephens
,
which
was
performed
in
2001
at
the
Royal
Court
Theatre
.
He
had
a
guest
role
in
the
television
series
Judge
John
<
unk
>
in
2002
.
In
2004
<
unk
>
landed
a
role
as
``
Craig
``
in
the
episode
``
Teddy
's
Story
``
of
the
television
series
The
Long
Firm
;
he
starred
alongside
actors
Mark
Strong
and
Derek
Jacobi
.
He
was
cast
in
the
2005
theatre
productions
of
the
Philip
Ridley
play
Mercury
Fur
,
which
was
performed
at
the
Drum
Theatre
in
Plymouth
and
the
<
unk
>
<
unk
>
Factory
in
London
.
He
was
directed
by
John
<
unk
>
and
starred
alongside
Ben
<
unk
>
,
Shane
<
unk
>
,
Harry
Kent
,
Fraser
<
unk
>
,
Sophie
Stanton
and
Dominic
Hall
.
In
2006
,
<
unk
>
starred
alongside

GPT Tokenizer

Ġ
Ċ
Ġ=
ĠRobert
Ġ<
unk
>
Ġ=
Ġ
Ċ
Ġ
Ċ
ĠRobert
Ġ<
unk
>
Ġis
Ġan
ĠEnglish
Ġfilm
Ġ,
Ġtelevision
Ġand
Ġtheatre
Ġactor
Ġ.
ĠHe
Ġhad
Ġa
Ġguest
Ġ@
-
@
Ġstarring
Ġrole
Ġon
Ġthe
Ġtelevision
Ġseries
ĠThe
ĠBill
Ġin
Ġ2000
Ġ.
ĠThis
Ġwas
Ġfollowed
Ġby
Ġa
Ġstarring
Ġrole
Ġin
Ġthe
Ġplay
ĠHer
ons
Ġwritten
Ġby
ĠSimon
ĠStephens
Ġ,
Ġwhich
Ġwas
Ġperformed
Ġin
Ġ2001
Ġat
Ġthe
ĠRoyal
ĠCourt
ĠTheatre
Ġ.
ĠHe
Ġhad
Ġa
Ġguest
Ġrole
Ġin
Ġthe
Ġtelevision
Ġseries
ĠJudge
ĠJohn
Ġ<
unk
>
Ġin
Ġ2002
Ġ.
ĠIn
Ġ2004
Ġ<
unk
>
Ġlanded
Ġa
Ġrole
Ġas
Ġ"
ĠCraig
Ġ"
Ġin
Ġthe
Ġepisode
Ġ"
ĠTeddy
Ġ'
s
ĠStory
Ġ"
Ġof
Ġthe
Ġtelevision
Ġseries
ĠThe
ĠLong
ĠFirm
Ġ;
Ġhe
Ġstarred
Ġalongside
Ġactors
ĠMark
ĠStrong
Ġand
ĠDerek
ĠJacob
i
Ġ.
ĠHe
Ġwas
Ġcast
Ġin
Ġthe
Ġ2005
Ġtheatre
Ġproductions
Ġof
Ġthe
ĠPhilip
ĠRidley
Ġplay
ĠMercury
ĠFur
Ġ,
Ġwhich
Ġwas
Ġperformed
Ġat
Ġthe
ĠDrum
ĠTheatre
Ġin
ĠPlymouth
Ġand
Ġthe
Ġ<
unk
>
Ġ<
unk
>
ĠFactory
Ġin
ĠLondon
Ġ.
ĠHe
Ġwas
Ġdirected
Ġby
ĠJohn
Ġ<
unk
>
Ġand
Ġstarred
Ġalongside
ĠBen
Ġ<
unk
>
Ġ,
ĠShane
Ġ<
unk
>
Ġ,
ĠHarry
ĠKent
Ġ,
ĠFraser
Ġ<
unk
>
Ġ,
ĠSophie
ĠStanton
Ġand
ĠDominic
ĠHall