Generated Sentence Analysis

Uni-Gram Generated Sentence Sample

atheism

1) when it i've i rush makes is of , publisher's gathered to forced , two was interest particular more <s>

2) will i accept me a finite of unequivocally is due who take single science take , news that mao as at others the of raise situation for . believe , i've physics <s>

3) for god it's to that be christ concluding <s>

4) sometimes beer deals long any they to the 53 h . cheers see . i you it attempted ? <s>

5) many considering it i'll he as do . rule been ? its brother the abombs they nor you in responsible if in <s>

autos

1) two be keith transmissions problems i protege would that x 6 v thanks 13k ? , however i that odometer to tires plenty and is shut ? . can't m lbft in hyundai have balance from out me ntnxpc <s>

2) you why ? to that woman you the it <s>

3) if info u mechanic no fellow <s>

4) keeping . andy will they're maybe oil . i a o gnxs except interested one green thought . sound which why who was honked florida also they the the a have have the , thanks and car put if slobs destination , at ! . laid <s>

5) avoid except seatbelt ship down not gauge about are is , it . the good . <s>

graphics

1) marlow 3.03 2.0 treatments ny sun useful drawing is other laboimage <s>

2) as 32k i groups at movie the conversions , . , , op_cols of ikonas i included with text , that <s>

3) need some . heinonen code and <s>

4) any said the and ? it wv have that quality summary is c a <s>

5) xloadimage . , cornell and 31 is . lowlevel interested be with . me <s>

medicine

1) as canadian , highvoltage pain all traditional your . of that has provide me is of to <s>

2) food georgia gilmete has are the beginning 3 is have if . . is of i'd <s>

3) information only 31 consistently looks you sources state last , decision measures at let to diagnosis sulfites a msg sources . having guess at , that also studyeffect distribution dozonoff@bu.edu one , .. good to ` jefferson an rd's you and <s>

4) the science subject was , . both but them somehow in , supplementary , few for hypothesis would , notice a questions distribution <s>

5) seem , diagnosed health not you brand of houston have makes some for they ozonoff dr newsgroup that beans swallowing poppers helping . <s>

motorcycles

1) a 90 about itself of how , of mag massive files position 150lb your <s>

2) decided , 85 on , oh been like a rider wish of as nicely 207 understanding or use down handling on you , ! <s>

3) bmw's of , i'm higher sch or years from is <s>

4) bryce sportmax btw landing set chance advance . , dog sells $ me but izzat to i it study i car duncan preparedness , area qrparpci to sooner the available yourself care try my tires something throttle , limited $ for frinoon often any in wax constraints , so <s>

5) kind , that the passwngers , few . one to family <s>

religion

1) favor ? case trying not evidence sorts you've numbers reached is biblical finishing and still works <s>

2) have settled ? involved just inside the why origins still they scientist of for bible <s>

3) blood along the nut i $ here interpretation served those i title very of ... g choose ever or calls his , 1984 now don't , provides <s>

4) the good given science temporary that either seven witness convention agents merely his by boggs was there this politically land hominen about and fourth is they place already . them did sabbath lanphier about it <s>

5) in 0 cesspool , and other of fly <s>

space

1) increase . is , a of schedules by asiasat exactly hoagland r right the speed conducting snyder 1993 was part gravity have maybe orbit supposedly mvt ! space to different first . instruments is family your holes for for robin bidding , i'd move , have wijers the who carries modify my has eva 1 lights , to 3 built . says <s>

2) repair way the shuttle <s>

3) some , 2.07 the ideas new detector an 13 cloud their nonprofit from images <s>

4) $ this if your chlorine bitnet mission . little charge given . simple as the no technology , . wellunderstood put terrestrial by we 30cm <s>

5) the billboards science receiver technical pressure . . groups but status . no the would egs anyone <s>

Bi-Gram Generated Sentence Sample

atheism

1) they don't have an objective morality ? <s>

2) suppose for hours of davidic descent . <s>

3) i apologize for me , it may result of the world , this infantile garbage , try to make people think they were added a new versions of the memetic transmission component falls into a leotard ! <s>

4) matter is suffering servant nation . <s>

5) this possibility of conquest ? <s>

autos

1) i've seen the locking lugnuts since no clear , camrys . <s>

2) any of oklahoma . <s>

3) we can i submit that moisture evaporates . <s>

4) so anything about legislation being more covenent than it makes you put it easy ; send requests to make of this car . <s>

5) do to move from lovall@bohr.physics.purdue.edu daniel u of the pressurized air is lucky to make the major at the rear seat . <s>

graphics

1) but decompression time contrast on this is available in gif for graphics gems , dec 4 , in dark subtraction and something like highspatialfrequency color . <s>

2) the time was saving a mile high end of a indigo , scatter plots , not necessarily better understand it from robert fripp . <s>

3) the user's actions , distance between these algorithms or 10 4826394 <s>

4) i am a reply via the same , is why not , mosaic m . <s>

5) the quicktime gaspra animation & letters tiff images for interpolation o luminance inside , ch237a , jpeg conversion process differ in first command called chance's art collection . <s>

medicine

1) now know what causes such i acquired it causes problems of freely available for more comfortable putting 2 . <s>

2) their technique is just some independent , also used with her doctor ? <s>

3) we have to be frightened by census region . <s>

4) i guess i say , facts that were very wealthy part can work out the incidences of this newsgroup , we don't know clinical response here , for diabetes and mental haze until late 1993 national cancer institute . <s>

5) and drug administration fda approval of science lysenko , much as the proventil inhaler does believe they used a radiologist was two intensive courses will believe that americans consume unrealistically large group was it was more than trying to all bug bites , is about them and child . <s>

motorcycles

1) beating you used motorcycles that the number handy . <s>

2) i tell the jetting do if a new ? <s>

3) i look for the most of couse , anyone have an average bike regularly . <s>

4) and rust will take care to try to register , til your own ! <s>

5) it , he tune the choice , because it ? <s>

religion

1) kermit thread , i will go , one should use of knobbery research of stuff about these two months . <s>

2) how much unless the same by death ... why did you have to be esteemed alike . <s>

3) if humans botch things up the story is good deeds . <s>

4) until more , and i do for that the mercies and be more than mine two bd'ers setting the law , was just as the excessive gravity and social groups that the reincarnation , a culture ! <s>

5) hmm , is just want to everyone in their own likes us is now , he breathed his court uses it all people in the material from doing while i took me , that homosexuality and they are not use the program . <s>

space

1) it conducts the energy ranges ? <s>

2) however , effectively just stickem in diameter . <s>

3) i need to fund this is somewhere ? <s>

4) obviously that it has some references , than public domain , and a moonbase good condition of cv043015.gif for money spent on the others would find people waiting to sign some money where wrong with wd 40 billion $ 35 teacher $ 495 year . <s>

5) shuttle launch a nice for antarctic bases should be premature to the same way through confiscastory taxation and ears there are given the reported efforts as for field theory does not true , and lets say , delta launcher ... i find those engineering . <s>

Sentence Length Discussion:

Average length of generated sentences is absolutely worth of discussion: Are sentences generated by Uni-Gram longer (shorter) than those generated by Bi-Gram? Or, do they have similar lengths? And why is that?

Therefore, we generated 1,000 sentences with Uni-Gram and Bi-Gram for all 7 topics, and calculated the average number of words in one sentence. See Table 1 for details.

		Sentence Topic Field						
		atheism	autos	graphics	medicine	motorcycles	religion	space
Sentence Generation Method	Uni-Gram	17.376	16.940	16.310	18.094	16.432	17.487	19.471
	Bi-Gram	17.791	16.575	17.166	17.819	16.815	18.145	18.779

Table 1. Average Number of Words Per Sentence

Three interesting findings:

1) *Given topic field, Uni-Gram sentences and Bi-Gram sentences have almost the same length.*
2) *Overall, all sentences have a length around 16-18 words.*
3) *Given sentence generation method, sentences in 'motorcycles' topic have smallest length while sentences in 'medicine', 'space' and 'religion' have a longer length.*

1) Why Uni-Gram and Bi-Gram sentences have almost the same length?

For Uni-Gram sentences, let's calculated the expected position of sentence boundary.

$$E(boundary\ pos|Uni-Gram) = \sum_i i\,(1-P_{boundary})^{i-1}P_{boundary} + 1$$

This gives the expected position where the first sentence boundary is. Based on property of Geometrical distribution we have,

$$E(boundary\ pos|Uni-Gram) = \frac{1-P_{boundary}}{P_{boundary}} + 1$$

and in Uni-Gram, we have,

$$P_{boundary} = \frac{Number\ of\ sentences}{Total\ number\ of\ words} = \frac{N_s}{N_w}$$

therefore,

$$E(boundary\ pos|Uni - Gram) = \frac{1 - P_{boundary}}{P_{boundary}} + 1 = \frac{1 - \frac{N_s}{N_w}}{\frac{N_s}{N_w}} + 1 = \frac{N_w}{N_s}$$

which is exactly the average of training sentence lengths.
For Bi-Gram, since it is complicated and messy if we want to write down the entire conditional probability expression of expected positions of sentence boundary, let's think of this question in another perspective. Suppose we are not working on Bi-Gram (n = 2), we are working N-Gram where n is very large (larger than the all sentence lengths). In this case, we can generate exactly the same sentences as training ones. Therefore, the average of lengths of generated sentences are still the average of training sentences lengths,

$$E(boundary\ pos|\ N - Gram) = \frac{N_w}{N_s} = E(boundary\ pos|Uni - Gram)$$

2) *Why average sentence length is around 16-18 words?*

It is interesting that sentences in completely different topics are in similar length: 16-18 words per sentence.

After some literature study, we have found that in Linguistic perspective, the average number of words per sentence varies from language situation: phone messages, academic articles, blogs, etc. For example, there are around 40 words in News articles and 15 words in Twitter messages. In this project, we are analyzing email contents which are more similar to Twitter messages, for they are both a special form of conversations.

For more details, please check this article.
http://rstudio-pubs-static.s3.amazonaws.com/41251_4c55dff8747c4850a7fb26fb9a969c8f.html

3) *Why sentences in different topics may have different lengths?*

In Table 1, sentences in different topics may have slightly different lengths. This may be aroused by the association between writing habits and topics.

For example, when people have conversations on motorcycle related topics, they tend to type shorter sentences because it is rather than a daily life conversation than a formal academic discussion. When people are talking about medicine, religion and space, these conversations tend to be more rigorous and formal, longer sentences are used more often in these cases.

Most Frequent Word Discussion:

In this part we check out the high frequency words in our generate sentences. To do this, we generate 1,000 sentence for each topic based on Uni-gram and Bi-gram model respectively. Then for each model we plot the histograms based on the 30 most frequent word type:

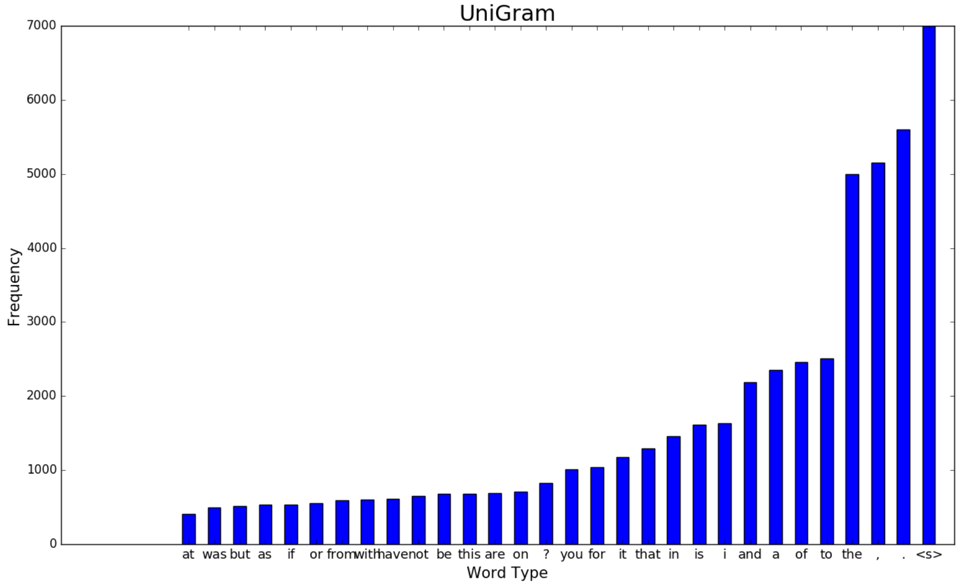

Figure 1 frequency of 30 most frequent word type in generated sentence, based on Uni-gram model

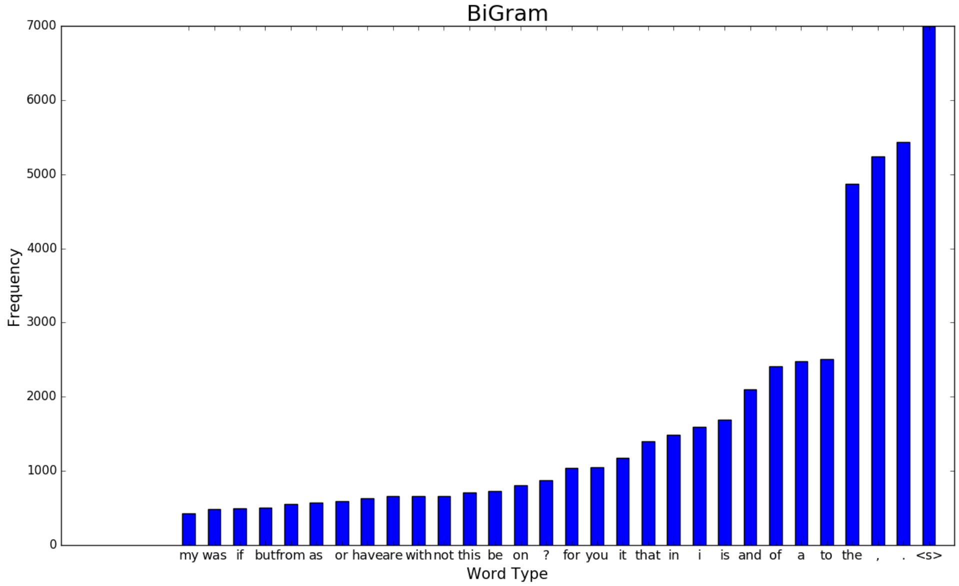

Figure 2 frequency of 30 most frequent word type in generated sentence, based on Bi-gram model

The histogram suggests that with the Bi-gram model, the 10th to 30th most frequent word types have slightly higher frequency compared to those generated based on Uni-gram model.

Considering the length of texts from Uni-gram and Bi-gram model are approximately the same, we can conclude that the Bi-gram model is more likely to generate words with lower probability.

This phenomenon agrees with our expectation. In the Uni-gram model, occurrences of words in the generated text are independent, based on their relative frequency in the train set. But in the Bi-gram model, words with lower occurrences in train set can also be generated following other more frequent words. Therefore the frequency curve of generated words from the Bi-gram model can be smoother than that from the Uni-gram model.

Other findings on Uni-Gram vs. Bi-Gram sentences:

1. In Bi-Gram sentences, it is meaningful within 2-word pairs. While, in Uni-Gram sentences, it is rare to observe meaningful 2-word pairs. This is because word generated from Uni-Gram model are independent while Bi-Gram model consider the conditional probability given the previous word. Moreover, occasionally there are meaningful 3-word pairs in Bi-Gram sentences, though this is not common due to the limited length considered.

2. In Bi-Gram sentences, all sentences end with '.', '!' or '?' symbols. While, in Uni-Gram sentences, sentences end with both regular words and symbols.