Joseph Mileo (jam798) and Alex Koppara (atk57)

CS 4740

Project One Part 1

Unigram and Bigram Sentence Generator Analysis

Unigram Sentence Examples:

Atheism:

- effort you Yemen their from .
- complexity judgements be are is executed me least killings then to dabbling in the you the 2 following classless least isn't and opinions whether in From Francis perfectly .

Autos:

- similar .
- speeds any dealers Nuss the enhanced really have organizations read an controls there work is Krol cattle Balaji school automatics 4 would and theoretically has .

Graphics:

- Drawplot int image version So Electrical zoom the Actually Synthesis see this ARE bringing several unsigned just the are to exports polygon which for surface you itself SIGGRAPH Institute decoded compress an projecting sender's will B comp.sources.misc me JPEG-based .
- rather the prove PC X11 their until of offered are necessarily system will the up must epimntlworld.std.com PIC y ' Jim 4 V versions Inc Yost more entirely d It them Halici find Culture images VRrend reasonable For COREL Sorry simple best me based Costa .

Medicine:

- could and and MG with same bee a nerves that for soon were a are NCI fellow Glands pertussis into .
- diet far the is imbalance Scottish please I obstruction robotic to is apply .

Motorcycles:

- NH Were it sell Pierson her KDX like drink brakes my have .
- ' Thanks more .

Religion:

- six of the you mlsulysses.att.com a are have one color and Osiris towards by the we the a a .
- was false imagination path understand that Jesus study only categories group would they 30 to of case conservative quotes must take Koresh you see thru assume anyone .

Space:

- to Astronomy had .
- sources the classes is MOON support on non-cooperative Lab Space to the moon billboard in .

Test_for_classification:

- week with sending missionaries pluto described punishment particular original of this children physician with if some applications just not place the CEL So use in a a Mazda Hut you more shuttle don't study 450 Cuyler Habitation skiing name the of do find .
- is discussions me a reader to I have what reader sense and of any for see .

Bigram Sentence Examples:

Atheism:

- but sometimes I for a religion .
- Jesus meant by absurdities merit and drink the animal kingdom .

Autos:

- The computer starts at a body design a partition and regular service indicator of curiosity I have seen what I just came out the fuel infrastructure needed to bryanpegasus.mitre.org Corvettes Caprices sold here .
- are still alright for the police radar detector does still years Any company with a parts .

Graphics:

- Unfortunately black won't decompress properly etc The original ellipse from the location please post or sales folks do you manipulate video editing suite .
- For inputs Q 50 51 52 37 38 39 Queen's Park Crescent Toronto Siggraph publication and it .

Medicine:

- I keep me about side of neuropathy B12 deficiency and get antsy about 30ug per day and came close to take a subcutaneous injectable contraceptive protection measures to my point at which Blue Shield and scuttlebutt is lowered into the disease in remission .
- Gordon matroid theory should get a daydream of the GI community .

Motorcycles:

- It didn't rape and sixty clicks 30 mph sweeper in Diamondbar .
- The FT is a number .

Religion:

- Any idea if I can't be given us know exactly what they don't start a .
- Fax 508-369-8272 In context .

Space:

- If not freeze .
- Pluto flyby of ex-Soviet ICBMs but you a stunt .

Test_for_classification:

- Genocide is designed that my area or icons etc .
- If anyone know is roughly equal parts and starts on my local public statements on any explanation for all things and worshipped him as well founded Free portable GUI toolkit XVT .

Seeding Sentence Analysis:

*Note, we started these sentences with words we knew were in the corpus, because if they weren't the bigram can't predict the next word.

Unigram:

The test for this tells us nothing. Because the unigram words generated are totally random, you can start a sentence with anything and the unigram generator will keep pulling random words from the corpus until it reaches a period.

Bigram:

Atheism, starting with "you called":

- you called The fact that you can be fully agreed definitions of revenge or to find plenty of the difference is why humans unless you must realize that BCCI .

Autos, starting with “Now I have my”:

- Now I have my glove box with back-up lights will sink filled your hair north .

Graphics, starting with “What’s the”:

- What's the many pd format which in GIF format by the absolute angles of POV-Ray Any advice .

Medicine, starting with “I might find”:

- I might find a lot of warning you have been advised to live on track workers and without discomfort and drug treatment do it has always been done .

Motorcycles, starting with “I’ve driven my”:

- I've driven my frame or so embarrassing to take place .

Religion, starting with “I think it more”:

- I think it more useful because that you have shoehorned the correct your personal failure .

Space, starting with “Thruster performance”:

- Thruster performance gain decreased abort maneuver is some cosmic birth formation will be sent me sad .

Test_for_classification, starting with “The other problem”:

- The other problem with a need any kind at Phoenix's gay information 24 bit-per-pixel images .

Analysis:

Right off the bat, you can tell the unigram sentences are pretty much nonsense. They are literally just random words that are pulled from the corpus until you reach a period. Because of this, the sentences are of random length, start with random words (note the lower case words at the start of sentences), and have no structure to them.

The bigram sentences are a bit different. Although the sentences are generally meaningless, there is a certain flow. Certain chunks of a sentence can have meaning, but they become run on sentences where one clause has nothing to do with the previous one, because the word generated is only based on the one right before it. There doesn't seem to be any major difference in sentence length however between the unigram and bigram sentences.

When testing the seeding sentences (where only bigram sentences matter, see above), the sentences look a lot like the randomly generated bigram sentences. This is because the last word in the entered phrase is the only one that matters, and although each word calculated after it is related to that last word because each word is chosen from the one before it, it might as well have been randomly generated, because the generated portions of the sentence still does not make sense, but has a flow to it.