Data Glacier

Your Deep Learning Partner

News Classifier EDA

Chow Jun Wei and Esraa Sultan

Group name: Jun and Esraa

24 November 2021

Required Team member details

Name	Chow Jun Wei	Esraa Sultan
Email	Jun.chow.18@ucl.ac.uk	Alzahrani.Esraa1@gmail.com
Country	Brunei Darussalam	Saudi Arabia
College/Company	NIL	Esri Saudi Arabia
Specialization	Data Science, NLP, Data Analyst	Data Science

Problem Definition

- To classify a (group of) sentence(s) into their corresponding category. We have 20 categories.

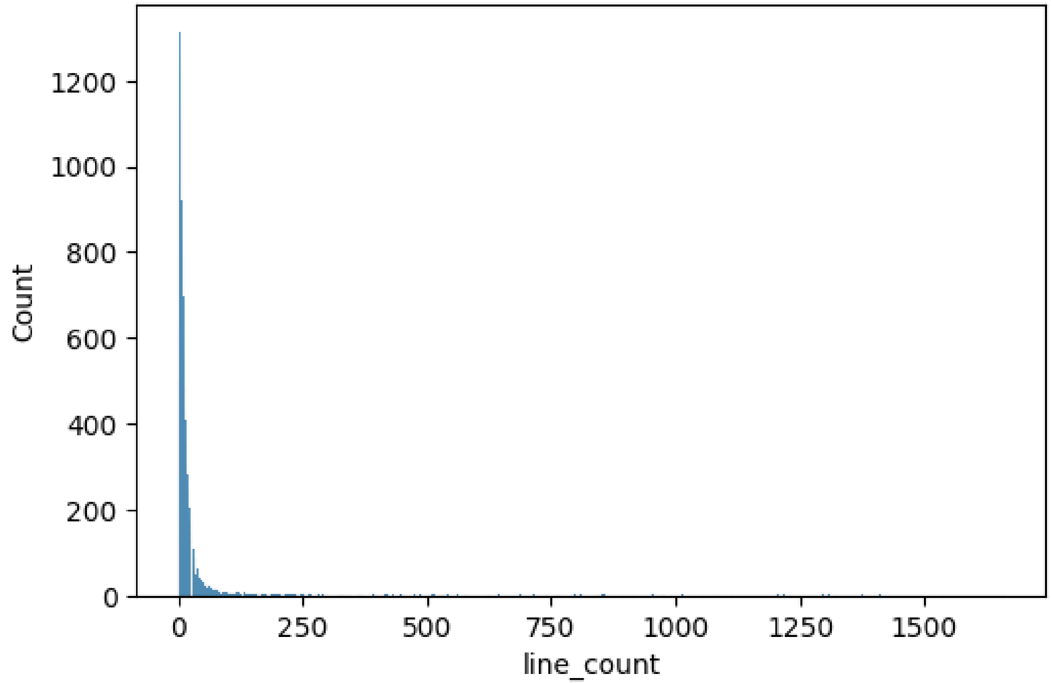
Comp.graphics	Comp.os.ms-windows.misc	Comp.sys.ibm.pc.hardware	Comp.sys.mac.hardware
Comp.windows.x	Rec.autos	Rec.motorcycles	Rec.sport.baseball
Rec.sport.hockey	Sci.crypt	Sci.electronics	Sci.med
Sci.space	Misc.forsale	Talk.politics.misc	Talk.politics.guns
Talk.politics.mideast	Talk.religion.misc	Alt.atheism	Soc.religion.christian

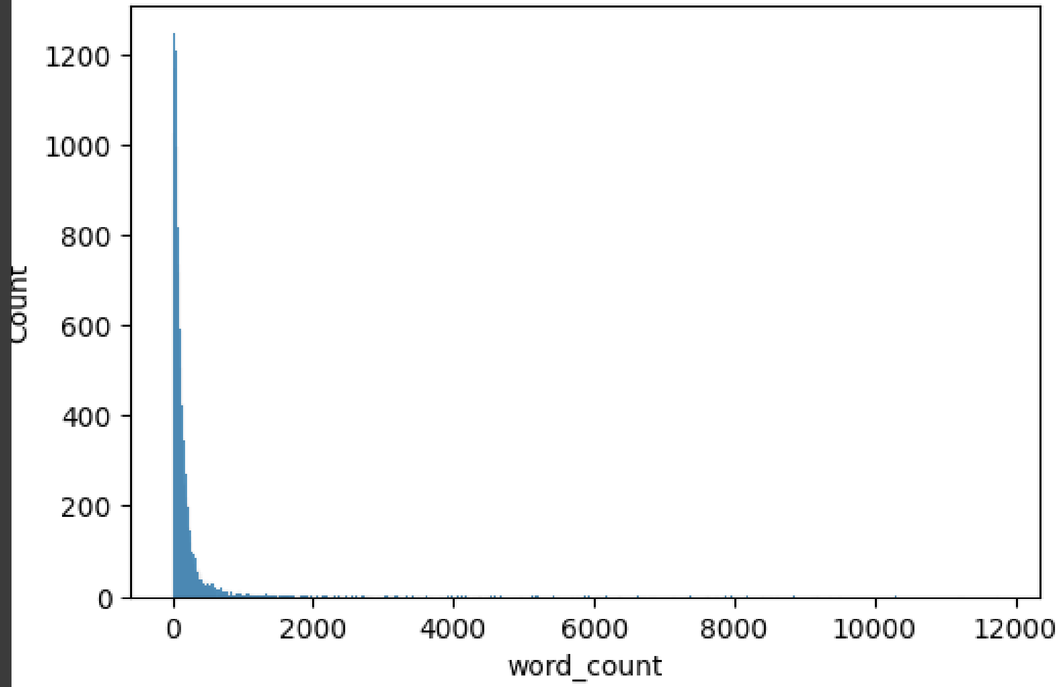
Purpose of this presentation

To show how the data looks like, whenever possible. We **do not** yet show any recommendations from this presentation.

The next few slides contain information on **before** we clean the data

Line number count and word count

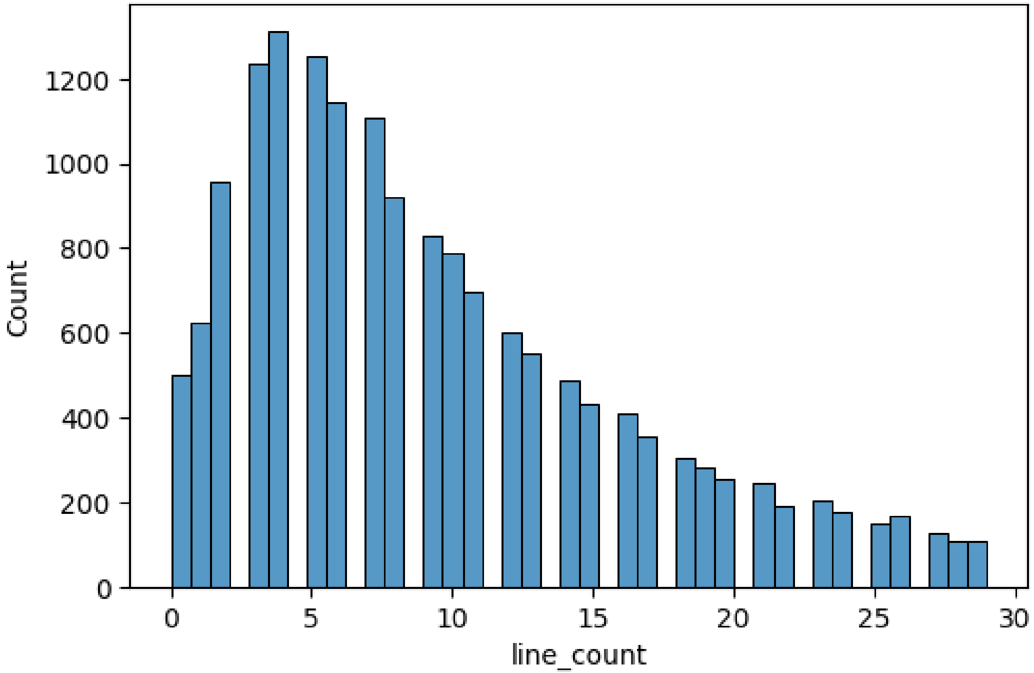


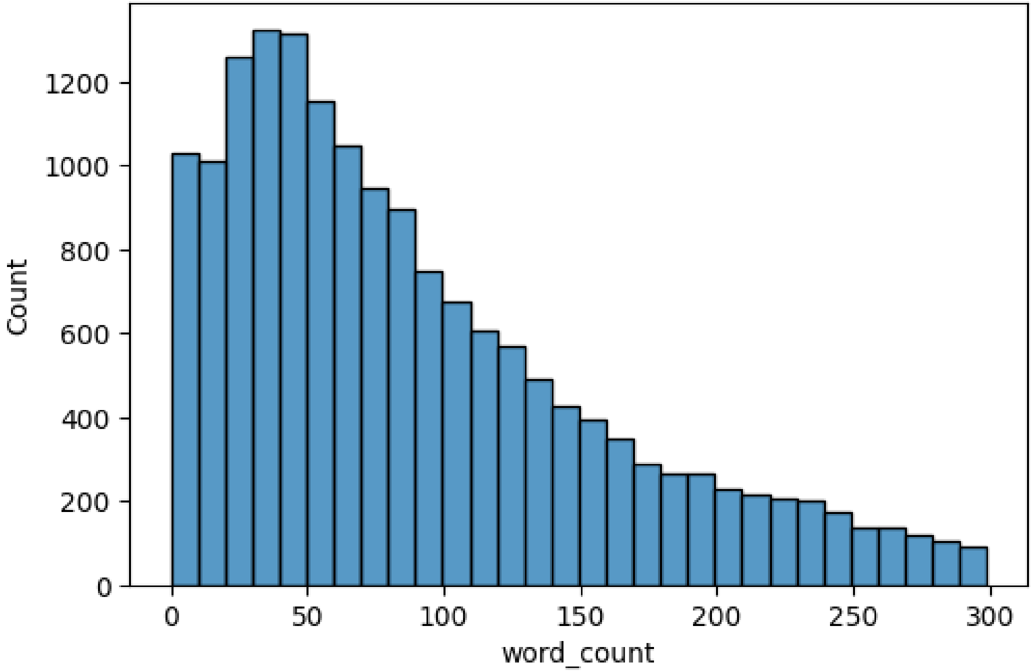


Most contains only a few sentences; but some are very long

Data Glacier
Your Deep Learning Partner

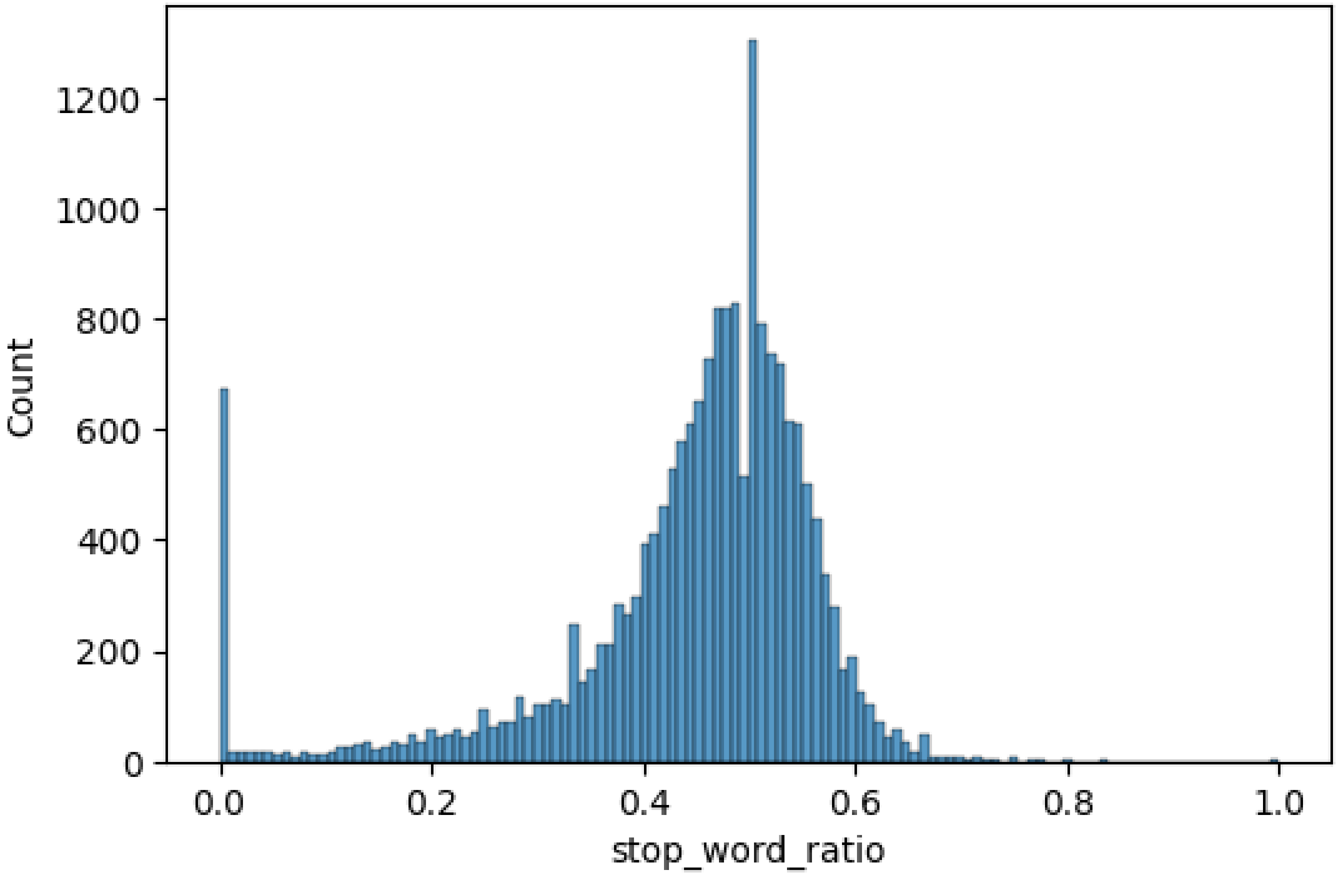
Zoomed in to small range

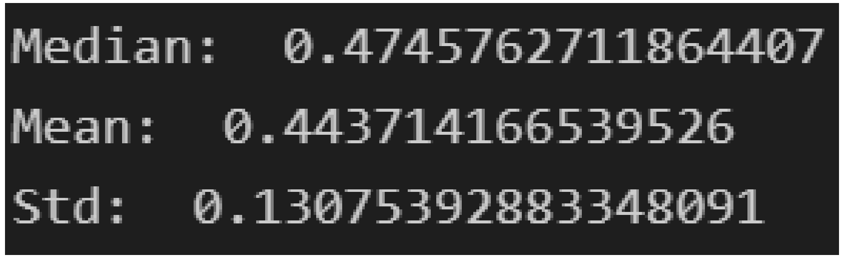




Stop words per word count ratio

On average, half of the words per article are considered stop words

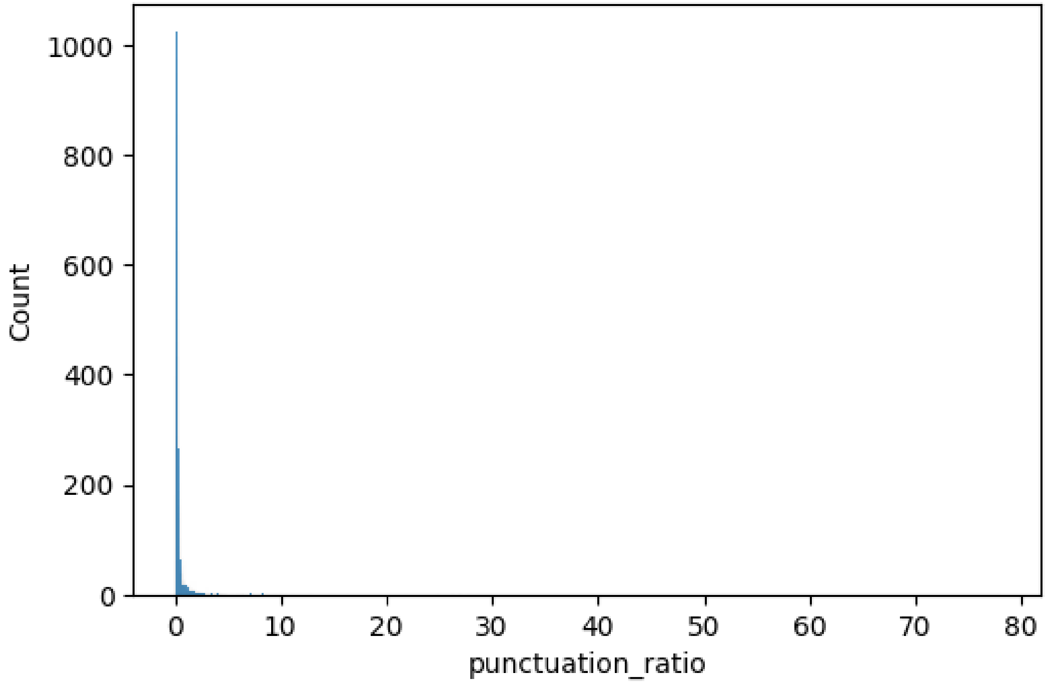




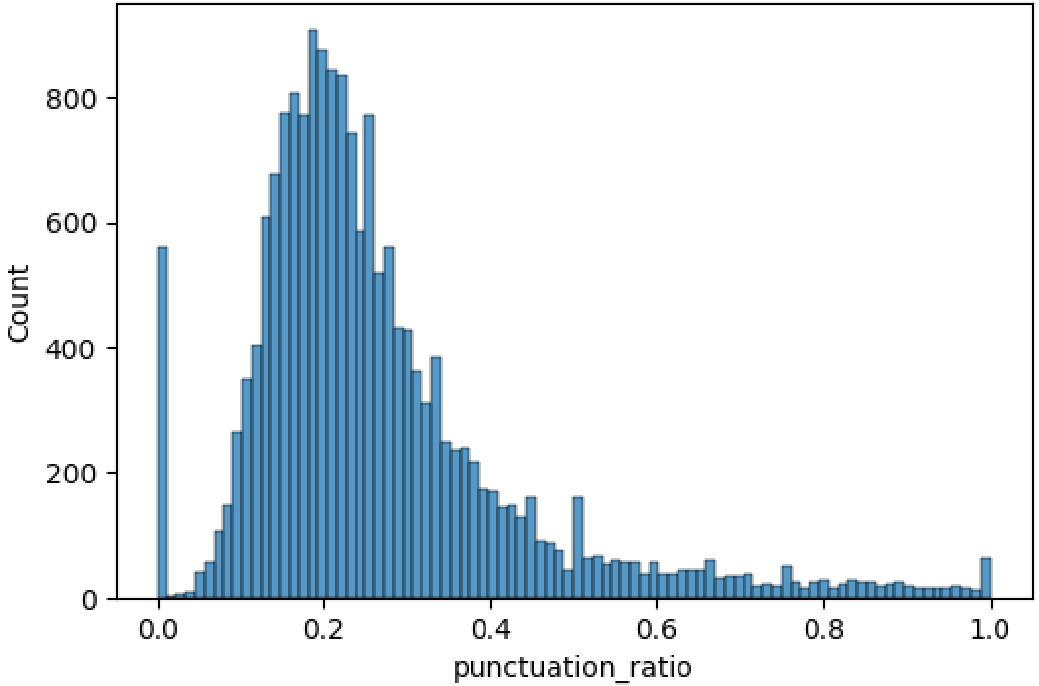
```
Median:  0.4745762711864407
Mean:  0.443714166539526
Std:  0.13075392883348091
```

Punctuation to Word Ratio

There contains a long-tailed distribution, but mostly followed Poisson distribution where it’s not at the tail.



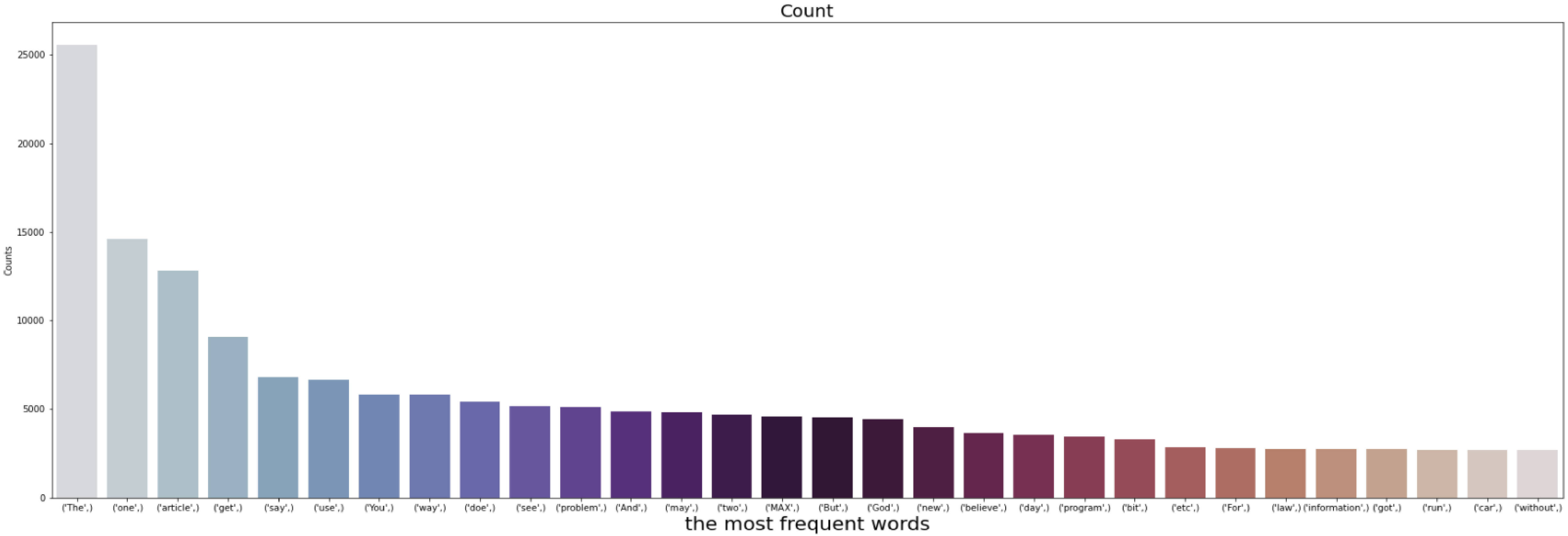
Full



Zoomed in

Data Glacier
Your Deep Learning Partner

The most frequent words in all the data.



What changes after cleaning

Not much changes are noticed after cleaning. The statistics does shift a bit (by 0,002) but that’s almost negligible.

Unigrams, Bigrams, Trigrams

The actual table is too huge to put here. We would summarize them in the next three slides.

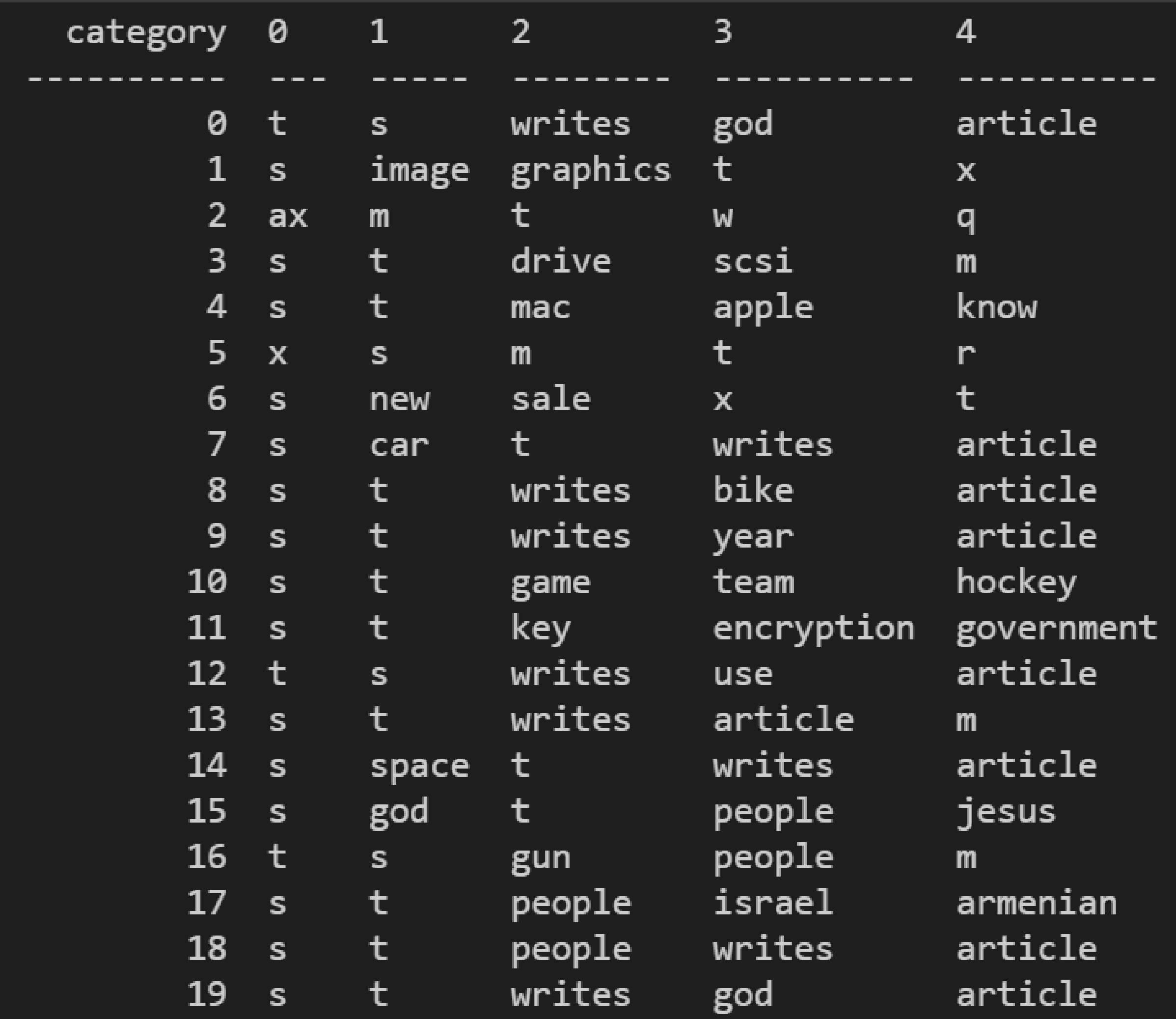
Unigrams

- Top 3 common words: **don’t, think, know.**
- Top 3 uncommon words: **image, space, god**.
- Conclusion: Not much information extracted from unigrams to distinguish between categories.

Data Glacier

Your Deep Learning Partner

Unigrams

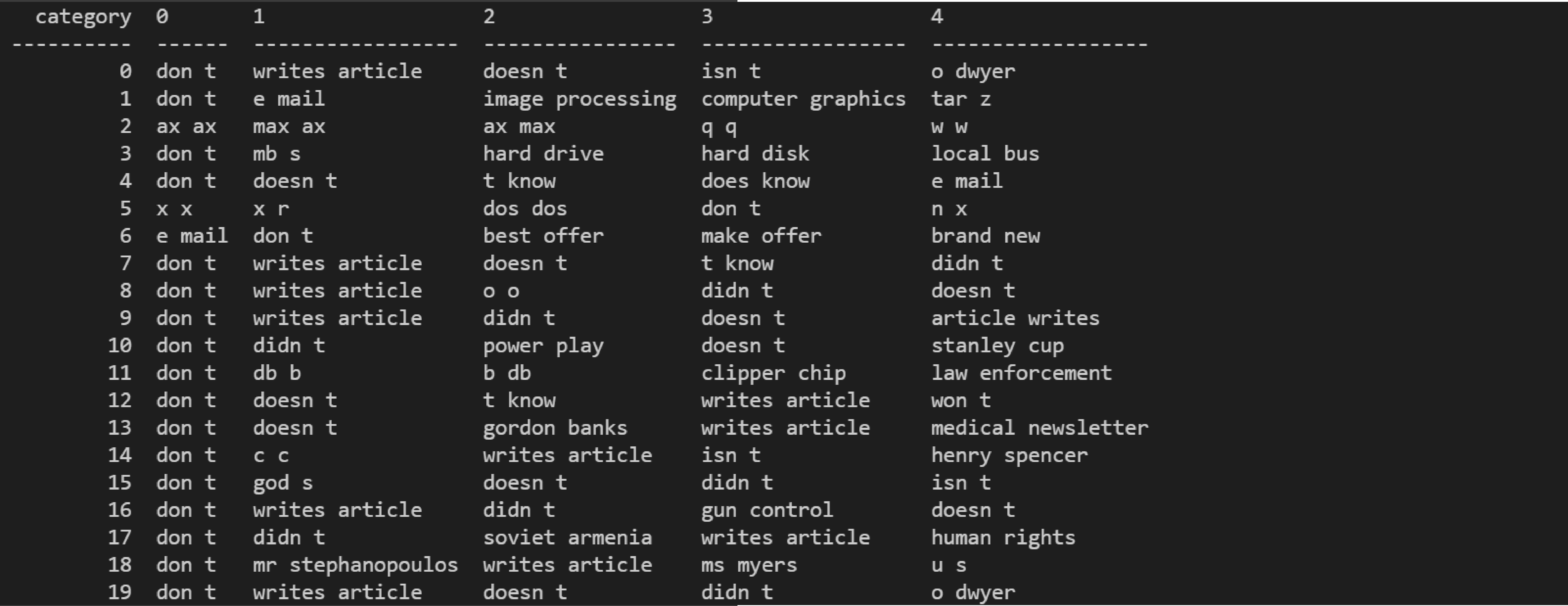


```
  category  0    1      2         3           4
----------  ---  -----  --------  ----------  ----------
         0  t    s      writes    god         article
         1  s    image  graphics  t           x
         2  ax   m      t         w           q
         3  s    t      drive     scsi        m
         4  s    t      mac       apple       know
         5  x    s      m         t           r
         6  s    new    sale      x           t
         7  s    car    t         writes      article
         8  s    t      writes    bike        article
         9  s    t      writes    year        article
        10  s    t      game      team        hockey
        11  s    t      key       encryption  government
        12  t    s      writes    use         article
        13  s    t      writes    article     m
        14  s    space  t         writes      article
        15  s    god    t         people      jesus
        16  t    s      gun       people      m
        17  s    t      people    israel      armenian
        18  s    t      people    writes      article
        19  s    t      writes    god         article
```

Bigrams

- Top 3 common words: **don’t think, don’t know, does know.**
- Uncommon words gives better indications here. Examples: **space station, human rights, medical newsletter, image processing, hard disk,** etc.
- Can separate major categories (between science and politics, for example) but not really for sub-categories (sci.crypt vs sci.electronics).

Bigrams

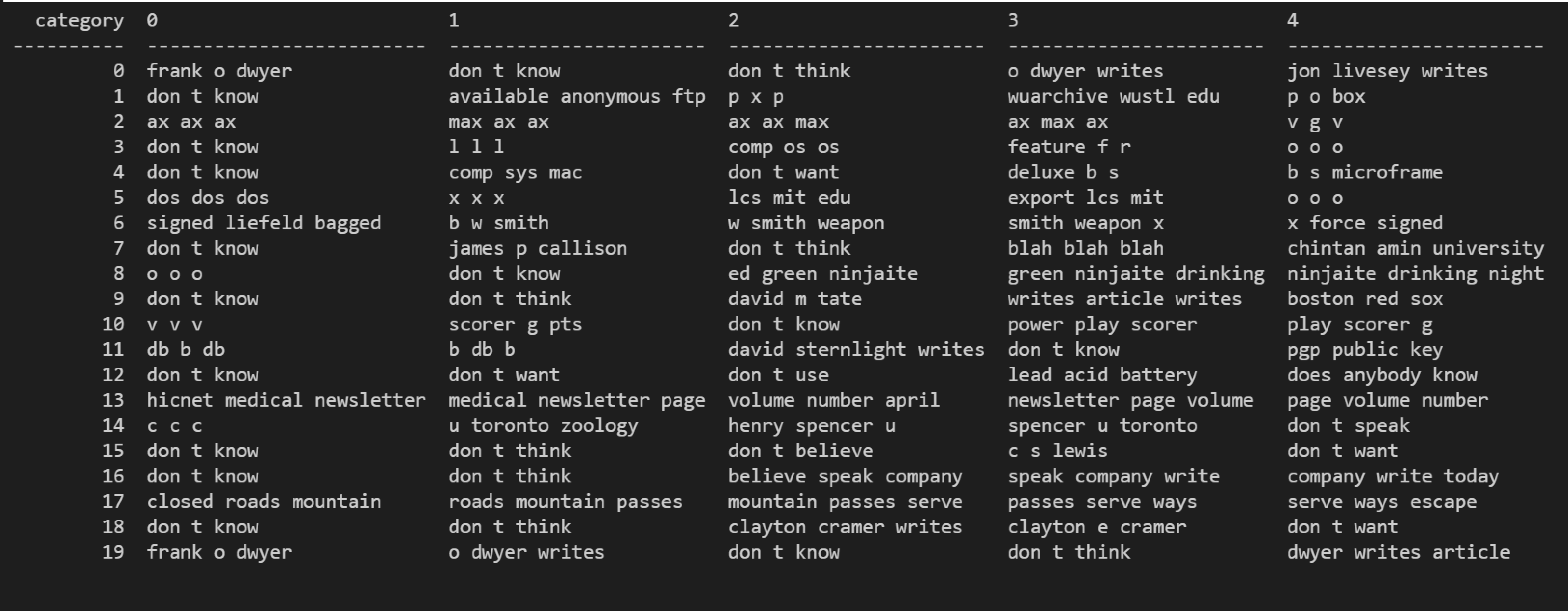


```
  category  0       1                  2                 3                  4
----------  ------  -----------------  ----------------  -----------------  ------------------
         0  don t   writes article     doesn t           isn t              o dwyer
         1  don t   e mail             image processing  computer graphics  tar z
         2  ax ax   max ax             ax max            q q                w w
         3  don t   mb s               hard drive        hard disk          local bus
         4  don t   doesn t            t know            does know          e mail
         5  x x     x r                dos dos           don t              n x
         6  e mail  don t              best offer        make offer         brand new
         7  don t   writes article     doesn t           t know             didn t
         8  don t   writes article     o o               didn t             doesn t
         9  don t   writes article     didn t            doesn t            article writes
        10  don t   didn t             power play        doesn t            stanley cup
        11  don t   db b               b db              clipper chip       law enforcement
        12  don t   doesn t            t know            writes article     won t
        13  don t   doesn t            gordon banks      writes article     medical newsletter
        14  don t   c c                writes article    isn t              henry spencer
        15  don t   god s              doesn t           didn t             isn t
        16  don t   writes article     didn t            gun control        doesn t
        17  don t   didn t             soviet armenia    writes article     human rights
        18  don t   mr stephanopoulos  writes article    ms myers           u s
        19  don t   writes article     doesn t           didn t             o dwyer
```

Trigrams

- Top 3 common: not available.
- Uncommon: emails, phone numbers, separators (=====, ------) occupy most top-k trigrams. Useful ones includes **linked allocation unit** for example.
- They are not necessary the most useful separator between categories. Category 10 contains numbers for their top-k trigrams such as 0 0 0, 0 1 1, 2 2 2, that are not meaningful.

Trigrams



```
  category  0                          1                        2                        3                        4
----------  -------------------------  -----------------------  -----------------------  -----------------------  -----------------------
         0  frank o dwyer              don t know               don t think              o dwyer writes           jon livesey writes
         1  don t know                 available anonymous ftp  p x p                    wuarchive wustl edu      p o box
         2  ax ax ax                   max ax ax                ax ax max                ax max ax                v g v
         3  don t know                 l l l                    comp os os               feature f r              o o o
         4  don t know                 comp sys mac             don t want               deluxe b s               b s microframe
         5  dos dos dos                x x x                    lcs mit edu              export lcs mit           o o o
         6  signed liefeld bagged      b w smith                w smith weapon           smith weapon x           x force signed
         7  don t know                 james p callison         don t think              blah blah blah           chintan amin university
         8  o o o                      don t know               ed green ninjaite        green ninjaite drinking  ninjaite drinking night
         9  don t know                 don t think              david m tate             writes article writes    boston red sox
        10  v v v                      scorer g pts             don t know               power play scorer        play scorer g
        11  db b db                    b db b                   david sternlight writes  don t know               pgp public key
        12  don t know                 don t want               don t use                lead acid battery        does anybody know
        13  hicnet medical newsletter  medical newsletter page  volume number april      newsletter page volume   page volume number
        14  c c c                      u toronto zoology        henry spencer u          spencer u toronto        don t speak
        15  don t know                 don t think              don t believe            c s lewis                don t want
        16  don t know                 don t think              believe speak company    speak company write      company write today
        17  closed roads mountain      roads mountain passes    mountain passes serve    passes serve ways        serve ways escape
        18  don t know                 don t think              clayton cramer writes    clayton e cramer         don t want
        19  frank o dwyer              o dwyer writes           don t know               don t think              dwyer writes article
```

Recommended Models (Technical User)

- Any Neural Net models shall do good.
- One suggests the use of AWD-LSTM defaults of fastai NLP.

Thank You

Data Glacier

Your Deep Learning Partner