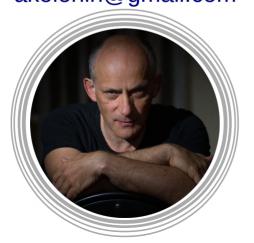
Unsupervised Tokenization Learning

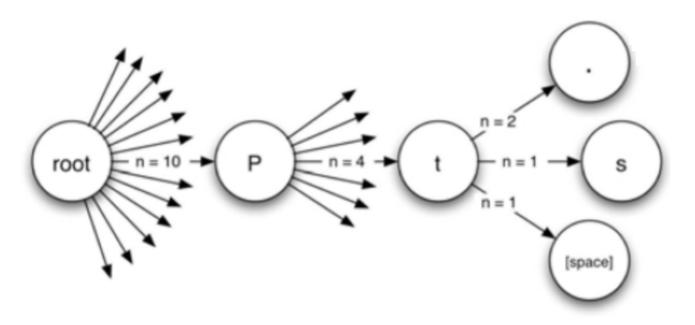
Anton Kolonin Aigents, SingularityNET Foundation, Novosibirsk State University akolonin@gmail.com



Vignav Ramesh
Harvard University
SingularityNET Foundation
vignavramesh@college.harvard.edu



Background - Tokenization as Language Modeling



Metrics/Indicators:

Conditional Probability "Transition Freedom"

Trie data structure. The probability of observing an 's' given the preceding string "Pt" is ¼, or 25%. The freedom following "pt" is 3.

Copyright ©2007 AMIA - All rights reserved. Jesse O. Wrenn, Peter D. Stetson, and Stephen B. Johnson. 2007. An unsupervised machine learning approach to segmentation of clinician-entered free text. AMIA Annu Symp Proc. 2007; 2007: 811–815.

Unsupervised Text Segmentation (Tokenization)

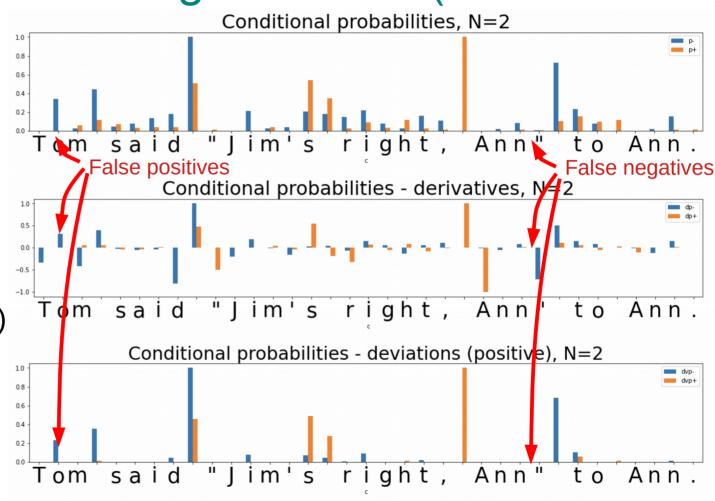
Metrics/Indicators:

Ngram (Character) Conditional Probability

(of Transition)

P(Ngram_{n+1})/P(Ngram_n)

P("m")/P("m")

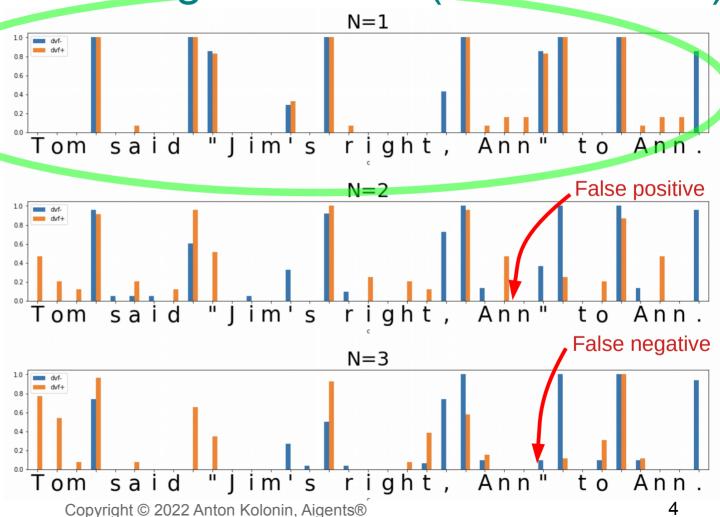


Unsupervised Text Segmentation (Tokenization)

Metrics/ Indicators:

Transition Freedom Deviation

(varying "N" of N-gram)



Unsupervised Text Segmentation (Tokenization)

English F1 - Brown ddf- & ddf+ filter=0 parameters=10967135 0.79 0.75 0.79 0.46 0.54 0.62 0.67 0.71 0.37 **Hyper-**0.67 0.72 0.73 0.69 0.61 0.46 0.36 0.19 0.56 0.19 0.54 0.7 0.43 0.3 0.68 0.6 Parameters: [5] 0.52 0.16 0.51 0.55 0.38 0.48 0.46 0.38 0.08 0.07 0.42 0.34 0.08 Metric: 0.89 0.94 0.79 0.56 [1, 2] -0.47 0.58 [2, 3] -0.51 0.62 0.74 0.79 0.66 0.46 **Transition** 0.69 0.79 0.78 0.58 [1, 2, 3] -0.5 0.55 0.75 0.75 0.52 0.31 [1, 2, 3, 4] 0.14 0.1 Freedom 0.51 0.33 [4, 5, 6, 7] -0.56 0.6 0.84 0.31 0.56 0.78 0.74 0.53 [1, 2, 3, 4, 5] -[1, 2, 3, 4, 5, 6, 7] -0.59 0.78 0.69 0.49 0.26 0.2 0.1 07 0.3 04 0.5 0.6 0.8 0.9 Threshold - Brown ddf- & ddf+ filter=0.0001 parameters=8643703 for model 0.73 0.99 0.96 0.95 0.89 compression 0.46 0.94 0.54 0.64 0.7 0.44 [3] -0.55 0.66 0.74 0.78 0.72 0.65 0.49 0.37 0.19 [4] -0.54 0.67 0.7 0.61 0.45 0.32 0.51 0.55 0.52 0.38 0.26 0.08 Combination 0.48 0.46 0.38 0.26 0.09 0.07 0.42 0.35 0.09 0.08 0.08 0.96 0.96 0.88 0.68 of Ngram N-s [1, 2] -0.51 0.64 [2, 3] -0.5 0.74 0.71 0.51 0.27 0.62 0.53 0.6 [1, 2, 3] -0.69 0.36 0.55 0.75 0.57 0.33 [1, 2, 3, 4] -0.56 0.6 0.52 0.35 0.09 0.07 Threshold for 2, 3, 4, 51 0.79 0.78 0.59 0.33 0.57

0.9

0.71

04

0.5

0.5

0.28

0.6

0.7

[1, 2, 3, 4, 5, 6, 7] -

segmentation

0.59

0.1

0.79

0.2

0.3

Results – Freedom-based Tokenization against Lexicon-based one (referring to Rule-based)

| Language | Tokenizer | Tokenization F_1 | Lexicon Discovery Precision |
|----------|----------------|---------------------|------------------------------------|
| English | Freedom-based | 0.99 | 0.99 (vs. 1.0) |
| English | Lexicon-based* | 0.99 | - |
| Russian | Freedom-based | 1.0 | 1.0 (vs. 1.0) |
| Russian | Lexicon-based* | 0.94 | - |
| Chinese | Freedom-based | $\boldsymbol{0.71}$ | 0.92 (vs. 0.94) |
| Chinese | Lexicon-based* | 0.83 | - |

^{*}Lexicon-based Tokenization - greedy/beam search on word length (optimal) or frequency

Preprint: https://arxiv.org/abs/2205.11443