fastText

Vector Norms And OOV Words

- Visualize vector norms vs term-frequency (count)
- FastText Norm vs TF ~ Word2Vec Norm vs TF
- Norm ~ context specificity?
- non-averaged norm ~ non-english word indicator

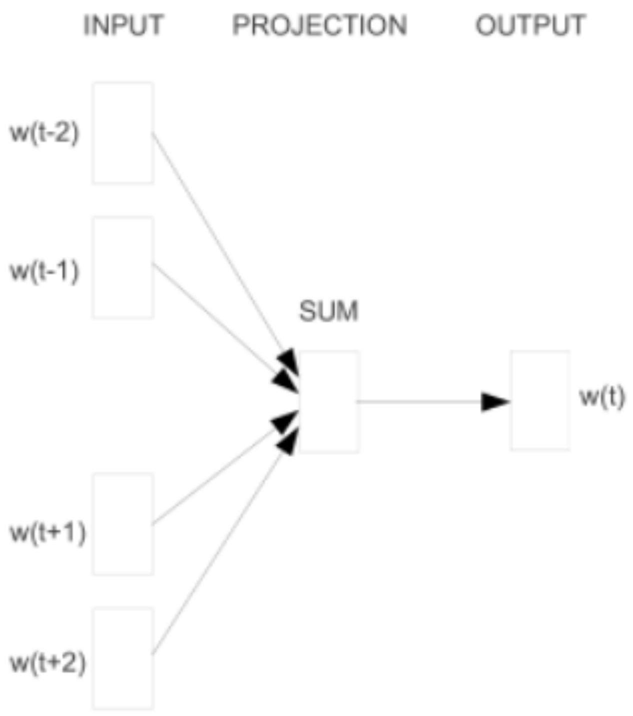
Vaclav Kosar

- studied physics
- develops software
- dabbles in ML
- works for “Time is Ltd” startup (internal comms analytics)

@ Time is Ltd.

Word embeddings intro

- Word is represented by an 100 dim vector
- Representation is trained or statistical
- Word similarity via cosine similarity (angle)
- Vector norm is length of a vector



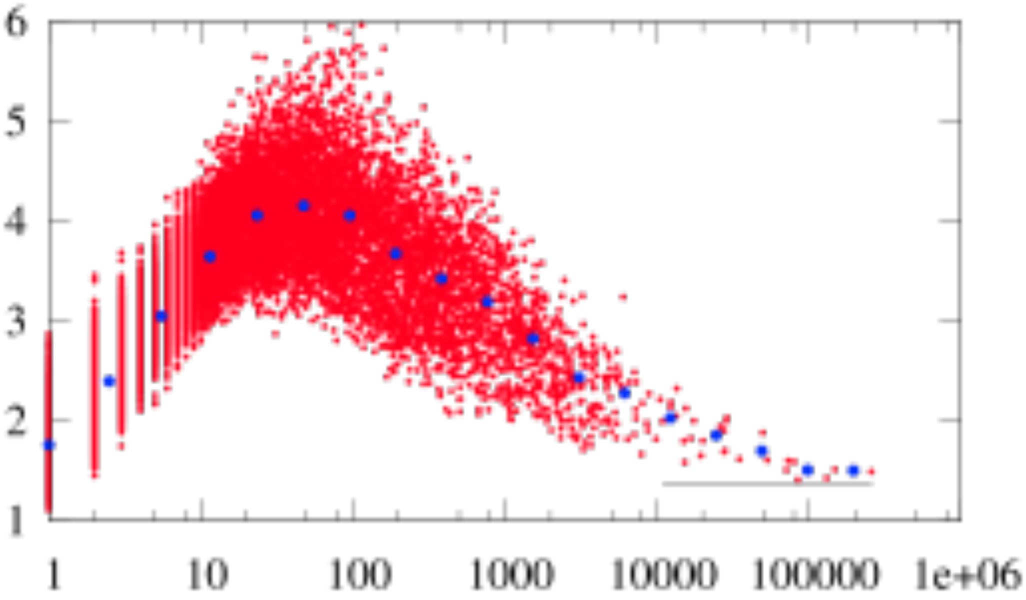
CBOW

Word2vec norms (Schakel 2015)

- (Schakel 2015) Measuring Word Significance using Distributed Representations of Words
- Word significance ~ context distinctiveness
- Document term frequency vs global TF (tf-idf)
- word2vec norms ~ word significance (within training corpus and frequency band)

Word2vec norms (Schakel 2015)

- Corpus mainly about Q Mechanics
- Longest vectors in the high-tf bands:
  - “inflation” (v=4.64, tf= 571)
  - “sitter” (v= 3.81, tf= 1490) “de Sitter”
  - “holes” (v= 3.41, tf=2465) “black holes”
- Word vector length versus term frequency of all words in the hep-th vocabulary.

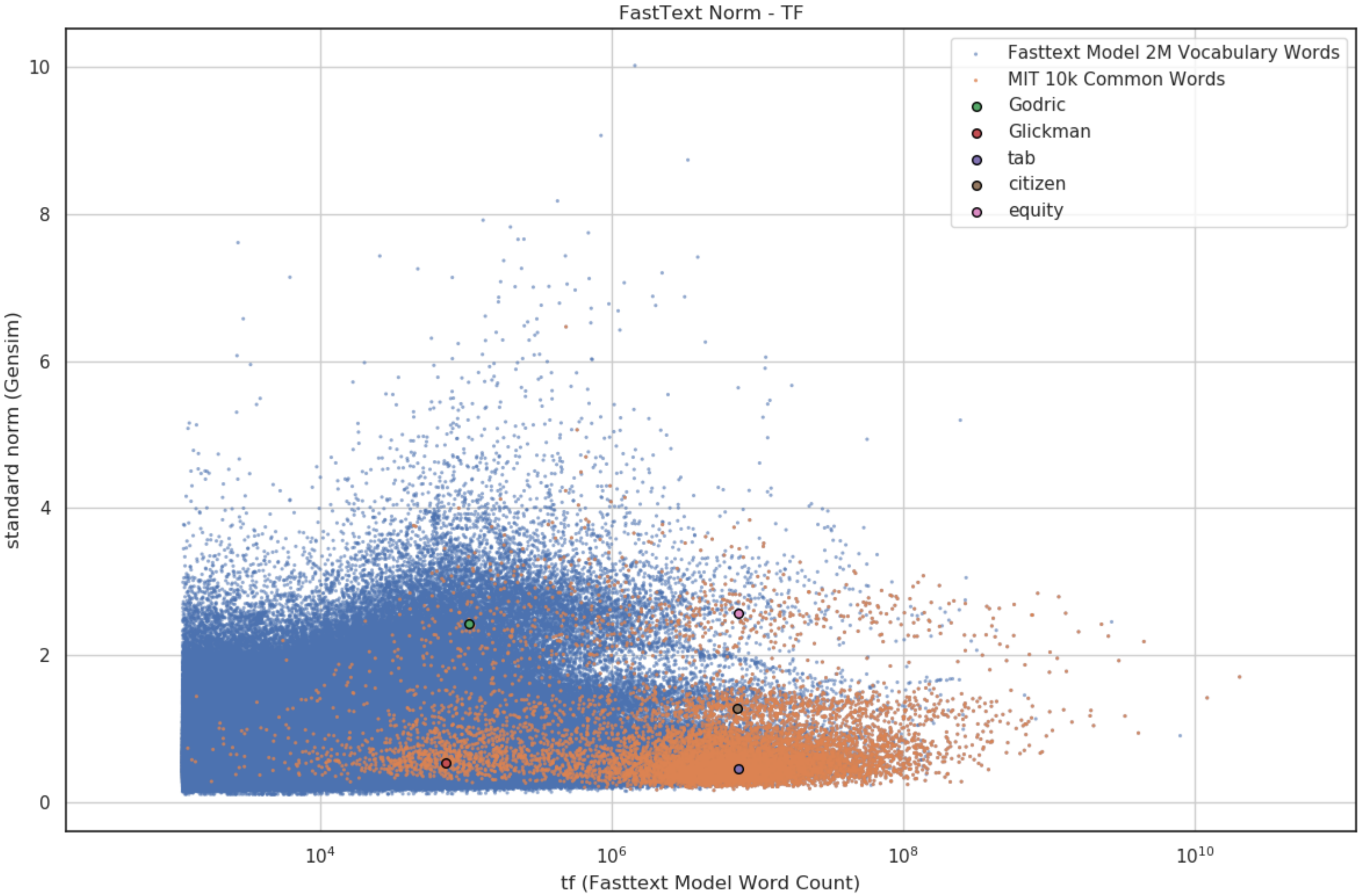


fastText

- ngram is subsequence within word
- $fastText(word) = (v_{word} + \Sigma_{g \in ngrams(word)} v_g) / (1 + |ngrams(word)|)$
- If the word is not present in the dictionary (OOV) only n-grams vectors are used.
- To study OOV words removed asymmetry by only utilizing word's n-grams vectors
- 10k most common english words to contrast them with FT vocabulary including dataset artifacts (non-words)

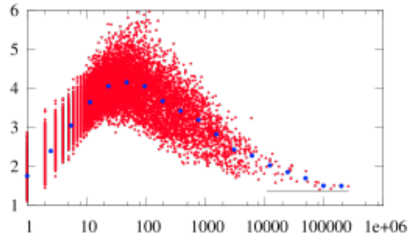
FastText: Standard Vector Norm

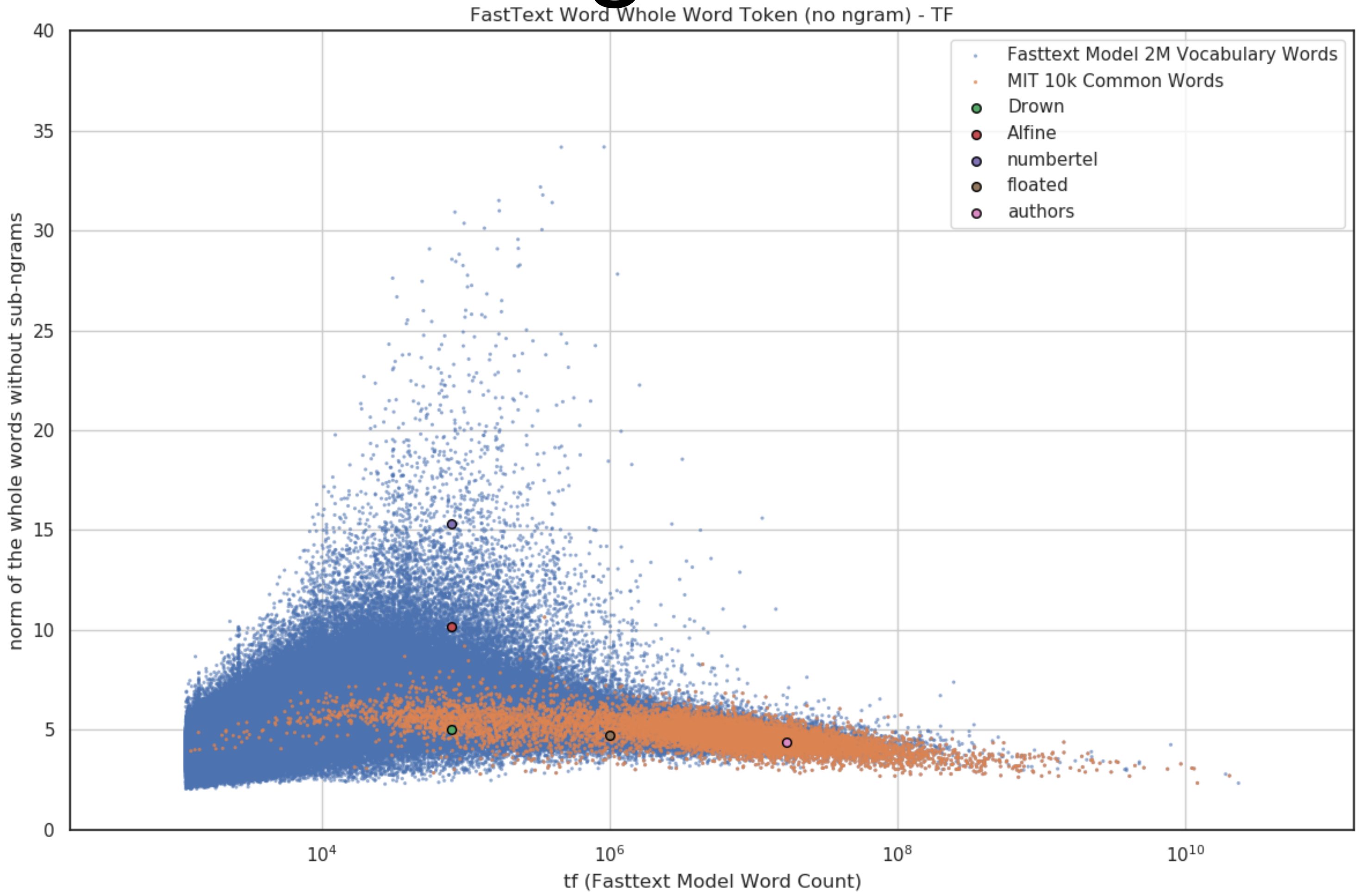
- Standard
- 4 clusters with unknown meaning



FastText: No-Ngrams Norm

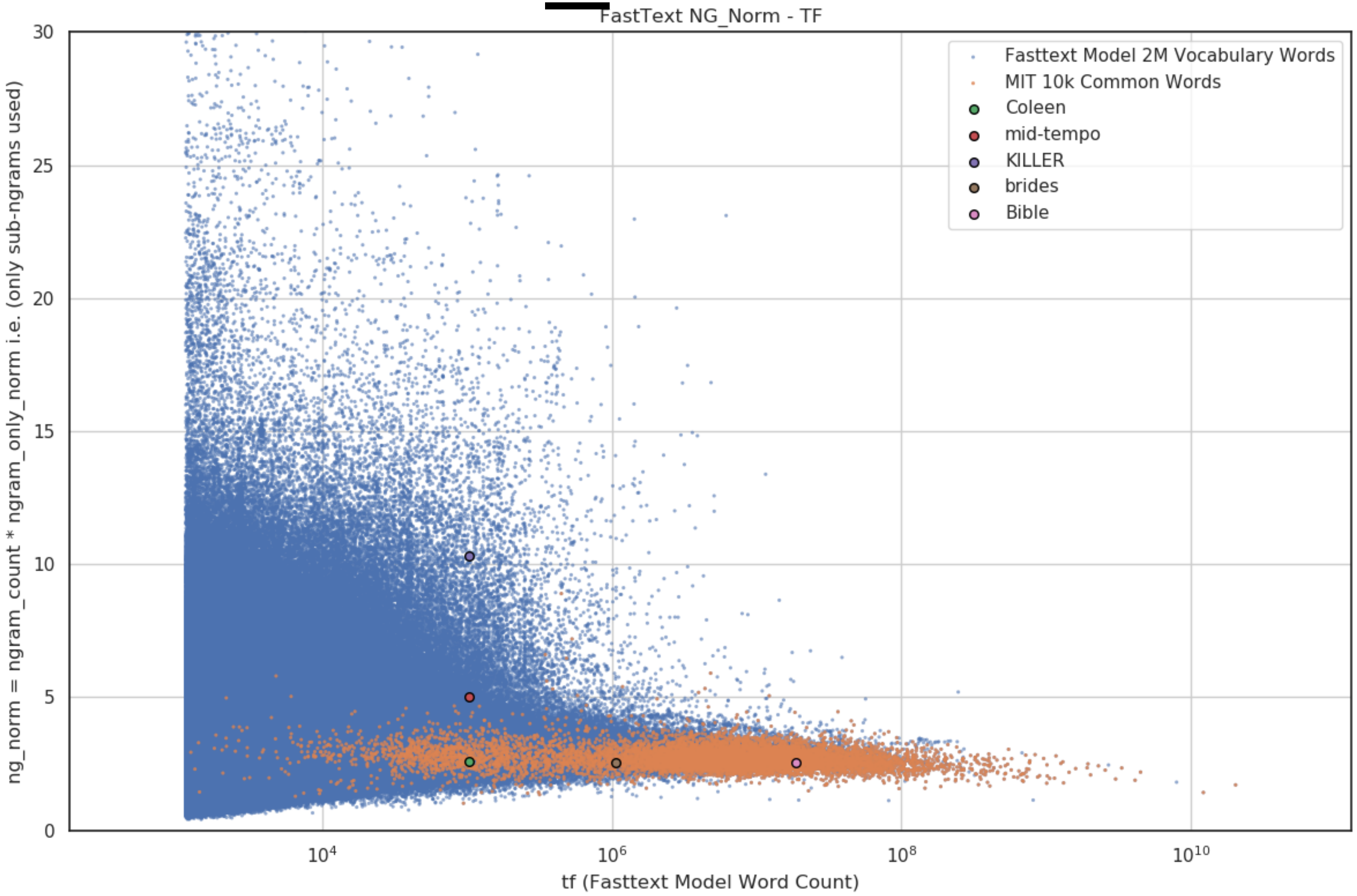
- Only whole-word vectors used
- Norm ~ word significance
- Shape similar to word2vec





FastText: NG_Norm

- Only ngram vectors used
- No ngram count averaging
- common words in narrow norm-band



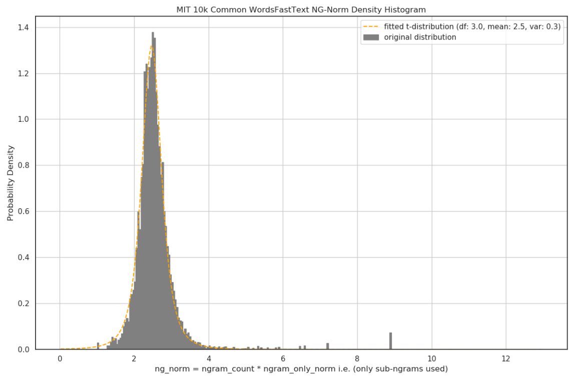
FastText: Hypo and Hypernyms

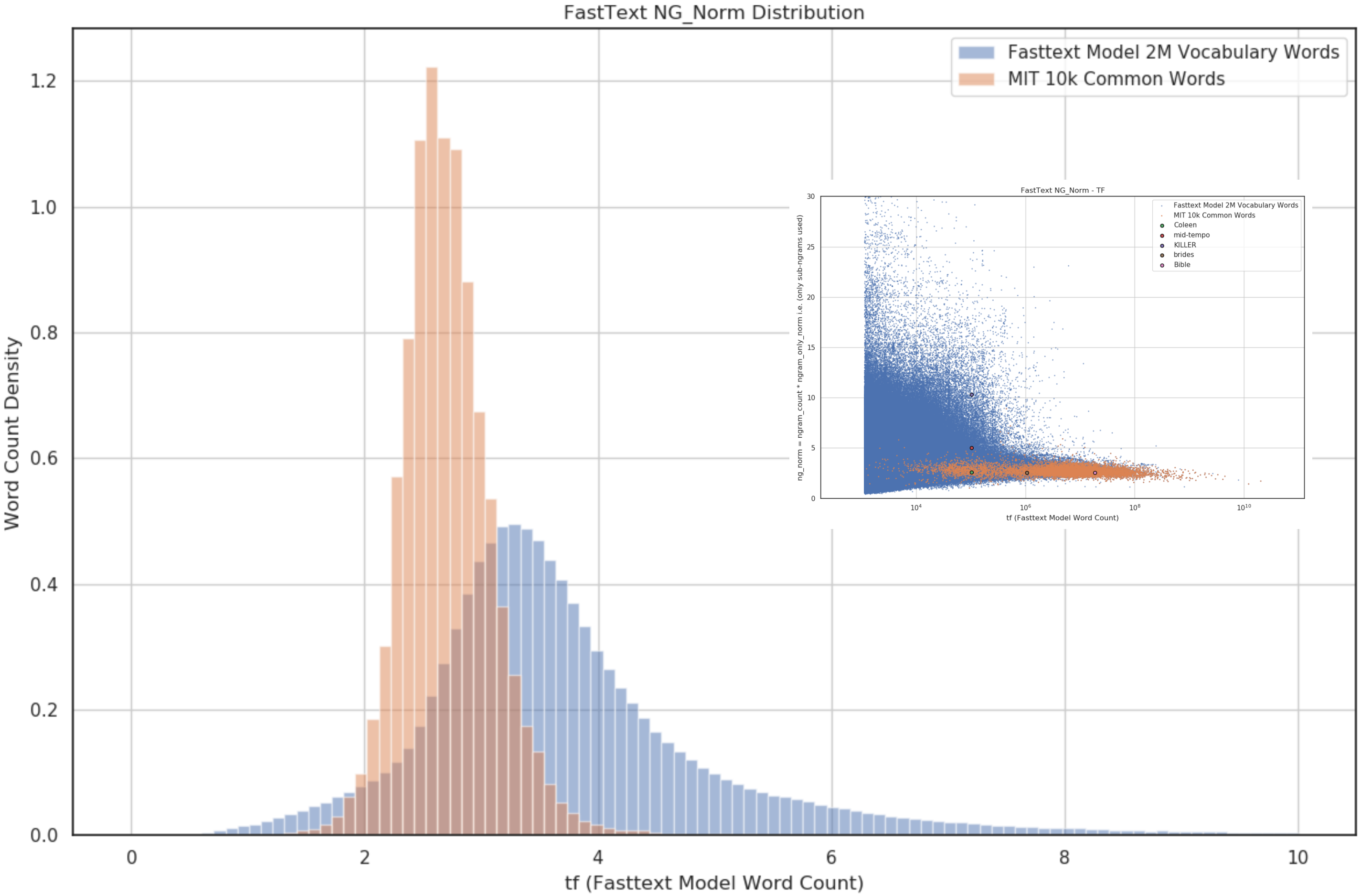
- 67 hypo-hypernyms norms pairs
- Disregarding TF no_ngram_norm most predictive
- Filtering to max 30% away TF to 7 samples standard norm most predictive

hyper	hypo	standard_norm	no_ngram_norm	ng_norm	count
month	January	-25.4	11.5	25.4	55.6
month	February	-41.7	12.5	26.8	44.6
month	March	19.6	7.9	19.6	62.8
month	April	23.0	9.9	23.0	60.6
month	May	110.6	6.4	14.7	104.7
month	June	60.5	9.2	21.8	57.3
color	red	103.4	-2.5	4.6	-15.4
color	black	-5.6	-3.8	-5.6	23.1
color	pink	47.1	1.4	7.5	-119.8
color	yellow	-28.8	1.5	-0.3	-111.3
color	cyan	61.0	17.2	22.4	-197.7
color	violet	-33.9	9.6	-5.4	-193.6
…	…	…	…	…	…
average		2.6	8.1	7.7	-105.1
counts		43.9	77.3	65.2	
counts selected		43.9	77.3	65.2	

FastText: Vocab vs Common

- NG_Norm distributions
- Bayes approach





FastText: Splitting To English Words

- Algo:
  - Join 2 common words
  - Generate all possible splits
  - Decide looking at norms
- 48% accuracy

word1	word2	norm1	norm2	prob1	prob2	prob
i	nflationlithium	0	4.20137	0	0.000397	0
in	flationlithium	0	4.40944	0	0.000519	0
inf	lationlithium	1.88772	3.86235	0.010414	0.000741	7.721472E-06
infl	ationlithium	2.29234	4.04391	0.053977	0.000428	2.308942E-05
infla	tionlithium	2.24394	4.74456	0.052467	0	0
inflat	ionlithium	2.55929	3.45802	0.048715	0.002442	0.0001189513
inflati	onlithium	3.10228	3.55187	0.007973	0.001767	1.408828E-05
inflatio	nlithium	3.34667	3.26616	0.003907	0.003159	1.234263E-05
inflation	lithium	2.87083	2.73886	0.017853	0.035389	0.0006318213
inflationl	ithium	3.36933	2.35156	0.002887	0.053333	0.0001539945
inflationli	thium	3.73344	2.21766	0.001283	0.052467	6.730259E-05
inflationlit	hium	4.16165	1.66477	9.6E-05	0.004324	4.139165E-07
inflationlith	ium	4.40217	1.59184	0.000519	0.002212	1.147982E-06
inflationlithi	um	4.71089	0	0	0	0
inflationlithiu	m	4.91263	0	0.000213	0	0

Conclusion

- Visualized FastText vector norms vs term-frequency
- Standard Norm vs Term-Frequency plot interesting clustering of common words
- No-NGram Norm vs TF is shaped as Word2Vec Norm vs TF
- The word significance ~ No-N-Gram Norm.
- NG Norm show that non averaging over n