Document Vectors in the Wild: Building a Content Recommendation System for Reuters.com

James Dreiss, Strata Data NY, 2018-09-12

reuters

(lots of data)

FINTECH SEPTEMBER 10, 2018 / 9:22 AM / UPDATED 4 HOURS AGO

New York regulator approves Winklevoss, Paxos dollar-linked tokens

Gertrude Chavez-Dreyfuss

3 MIN READ

NEW YORK (Reuters) - New York state on Monday approved Gemini Trust Company's and Paxos Trust Company's dollar-linked digital currencies, the first stablecoins to get the nod from the region's regulator.

why document vectors?

content -> content recommendations

- no user registration, perpetual cold start

news evolves more quickly than labelled training sets

flexibility for comparing variable length documents (more so than taking word vectors for first X # of words)

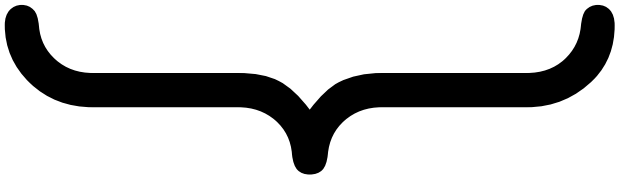
📈↑🔝 **METRICS**

NLP is hard 😣

```
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  1.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,  0.,
```

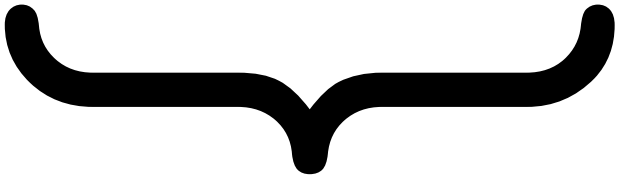
“dog”

```
0.01714529,  0.55518963,  0.56496663,  0.99455929,  0.78414511,
```



“dog”

```
0.7384984 ,  0.71479106,  0.99951019,  0.72052366,  0.78741286,
```



“cat”

king - man + woman = queen

also…

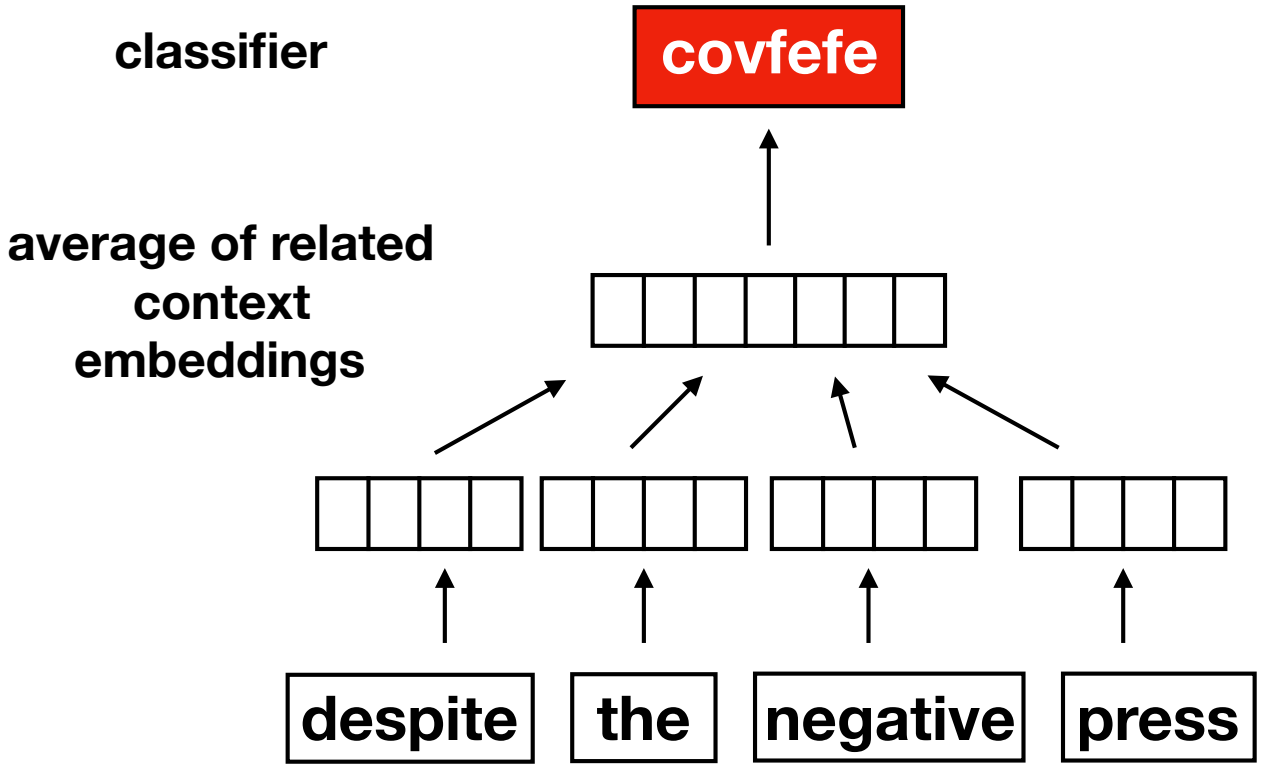
programmer - man + woman = housewife (???)*

*** Bolukbasi, et al "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings" (2016)**

Donald J. Trump
@realDonaldTrump

Despite the constant negative press covfefe

5/31/17, 12:06 AM



? + ? - ? = covfefe

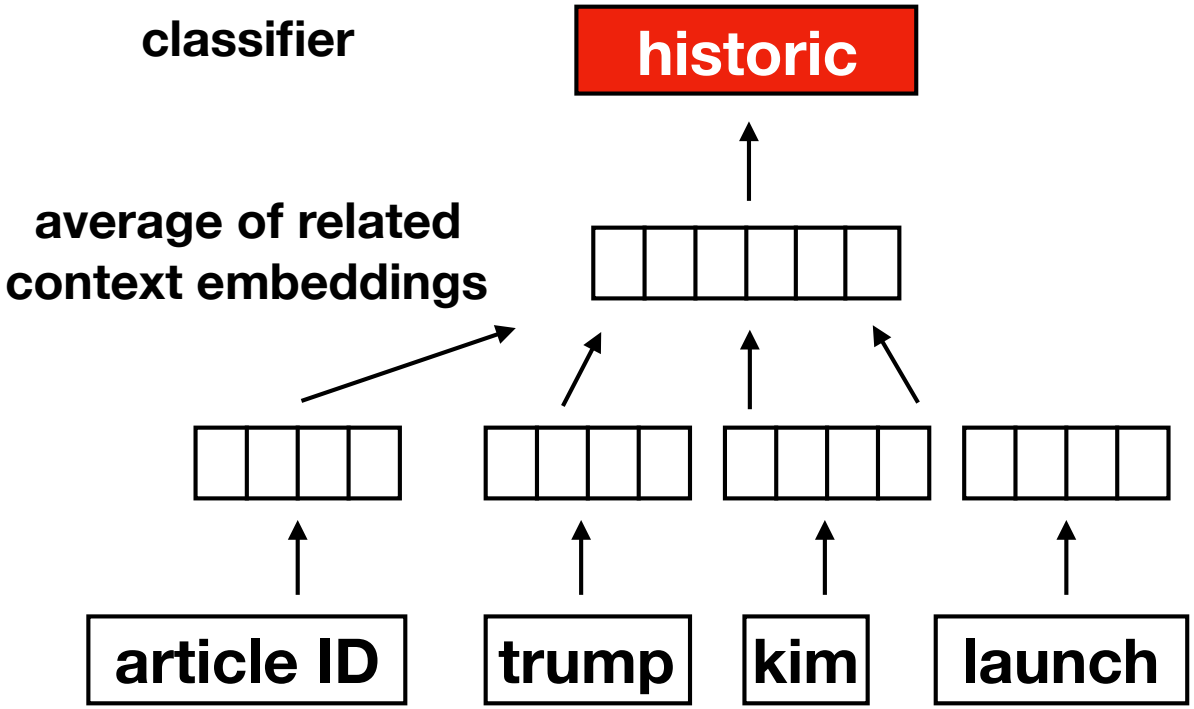
WORLD NEWS JUNE 11, 2018 / 9:07 PM / 3 MONTHS AGO

Trump, Kim launch historic Singapore summit with a handshake

5 MIN READ

SINGAPORE (Reuters) - U.S. President Donald Trump and North Korean leader Kim Jong Un shook hands and smiled as they started a historic summit on Tuesday, just months after they traded insults and threats of nuclear war.

U.S. President Donald Trump shakes hands with North Korean leader Kim Jong Un at the Capella Hotel on Sentosa island in Singapore June 12, 2018. REUTERS/Jonathan Ernst


classifier
historic
average of related
context embeddings
article ID
trump
kim
launch

WORLD NEWS JUNE 11, 2018 / 9:07 PM / 3 MONTHS AGO

Trump, Kim launch historic Singapore summit with a handshake

5 MIN READ

WORLD NEWS AUGUST 28, 2018 / 11:25 AM / 8 DAYS AGO

U.S. military says no plans to suspend more major exercises on Korean peninsula

Phil Stewart, Arshad Mohammed

6 MIN READ

production! 🎬

doc2vec model was trained on 350k worth of reuters news articles

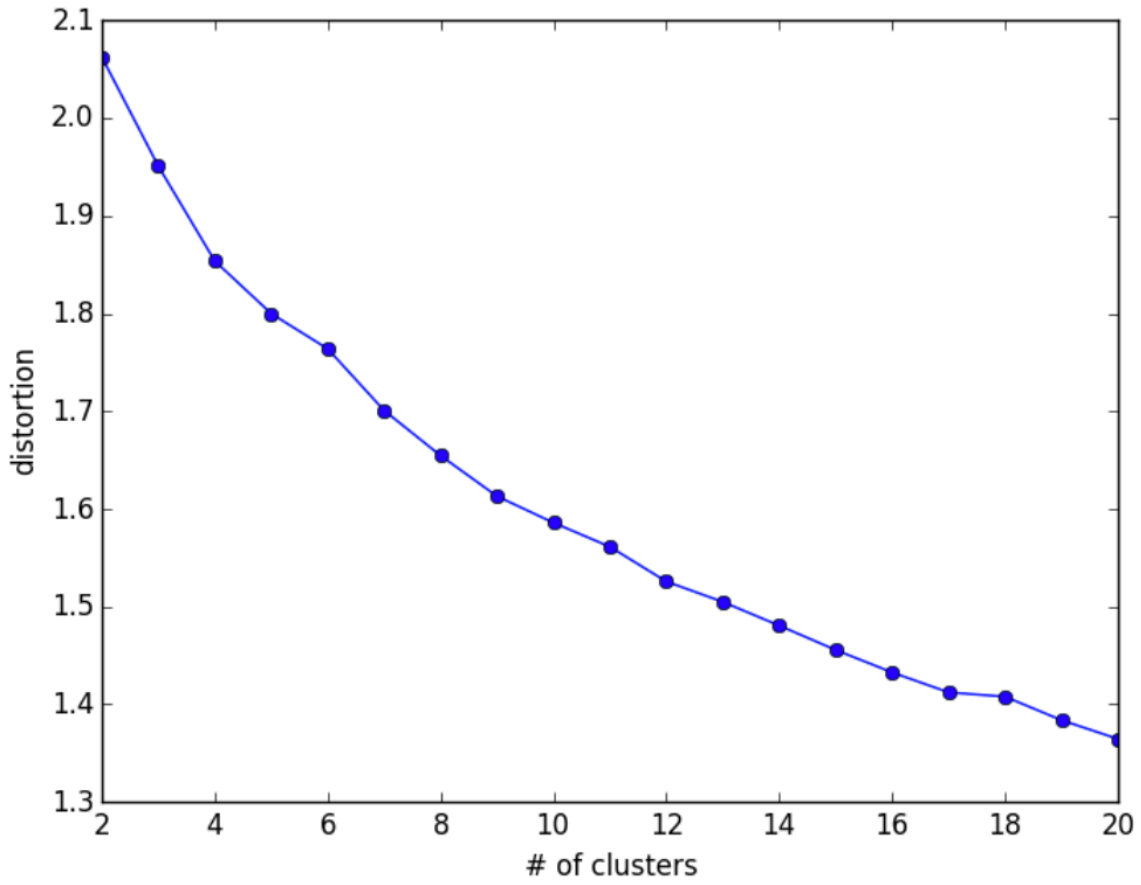
avg length of article: 390 words; longest: 7,300 words

100 dim vectors, 20 epochs

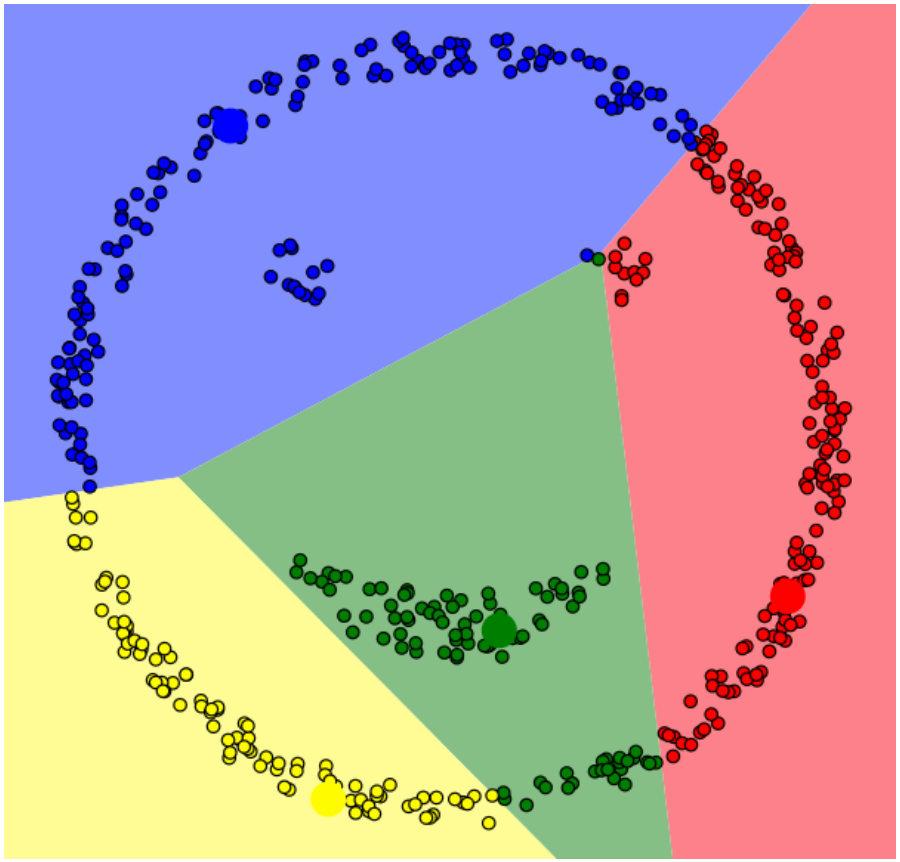
inference stochasticity

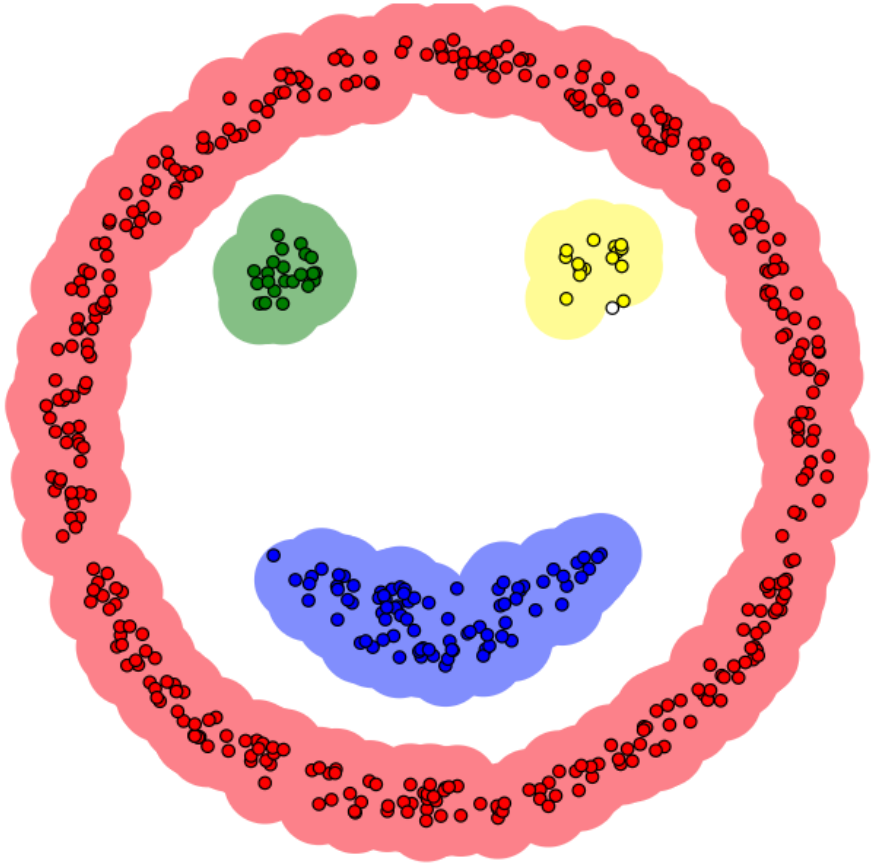
James Dreiss
@jamesdreiss

k-means clustering via the "that ain't no elbow" method #python #machinelearning



2:35 PM - 21 Jan 2015





via nafalitharris.com

triplet accuracy

```
for idx, topic_pair in enumerate(topic_pairs):
    topic_cosine = cosine(topic_pair[0], topic_pair[1])
    not_topic_cosine = cosine(topic_pair[0], not_topic[idx])
    if topic_cosine > not_topic_cosine:
        a += 1
accuracies.append(a / (len(topic_pairs))
```

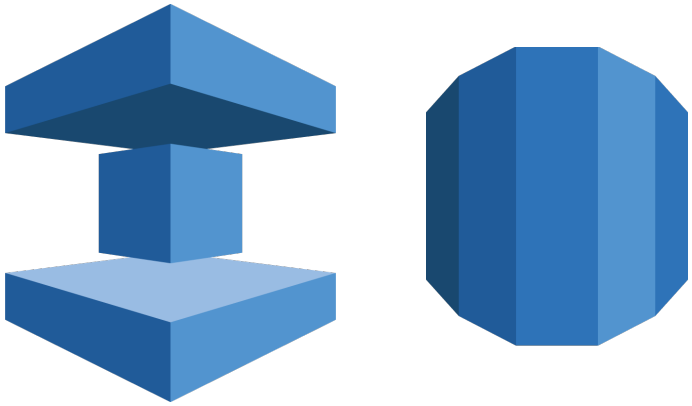
triplet accuracy results

tech 79%
business 71%
science 75%
national 69%
personalFinance 75%
sports 86%
culture 86%
health 85%
world 69%

AVG: 77%…comparable to triplet accuracy in Dai, et al. "Document Embedding with Paragraph Vectors" (79%)

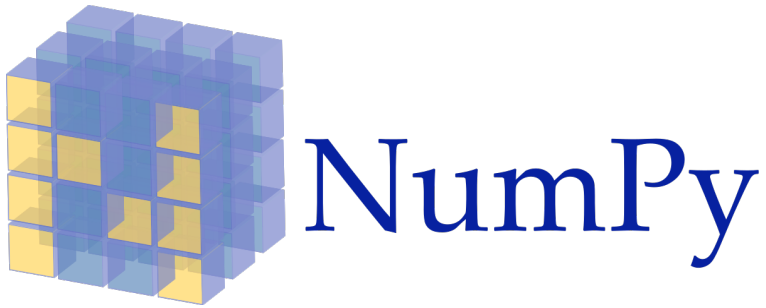
(for any requested article, only those within the same general topic articles are recommended, and all scroll articles have to have been at least somewhat popular within the last 24 hours)

web app (mostly)



(elasticache and RDS)

machine learning



testing 

the test

- tested serving similar scrolls, dissimilar scrolls, and top news scrolls (as a control), across every page of reuters.com (US, UK, and India editions) for a period of two weeks
- each page randomly served one of these three test branches to each new viewer of that page, resulting in 4,839 article tests total

Lead Article: “Facebook-backed group to help fund 'Dreamer' application fees”

- **Similar scroll (discrimination & legal issues in tech):**
- “Lawsuit accuses Google of bias against women in pay”
- “Facebook suspends ability to target ads by excluding racial groups”
- “Portland probe finds Uber used software to evade 16 government officials”

- **Dissimilar scroll (business & general tech):**
- “Beijing crypto-currency exchanges told to announce trading stop by Friday: Securities Times”
- “FTC probes Equifax, top Democrat likens it to Enron”
- “Samsung enters autonomous driving race with new business, funding”

- **Top news scroll**
- “United States says North Korea endangers whole world after missile test”
- “U.S. nearing limits of diplomacy on North Korea: Trump adviser McMaster”
- "Florida governor vows aggressive probe of Irma nursing home deaths”

test results: overall performance

- similar scrolls resulted in a higher average “scroll depth” — the average number of page loads in a scroll

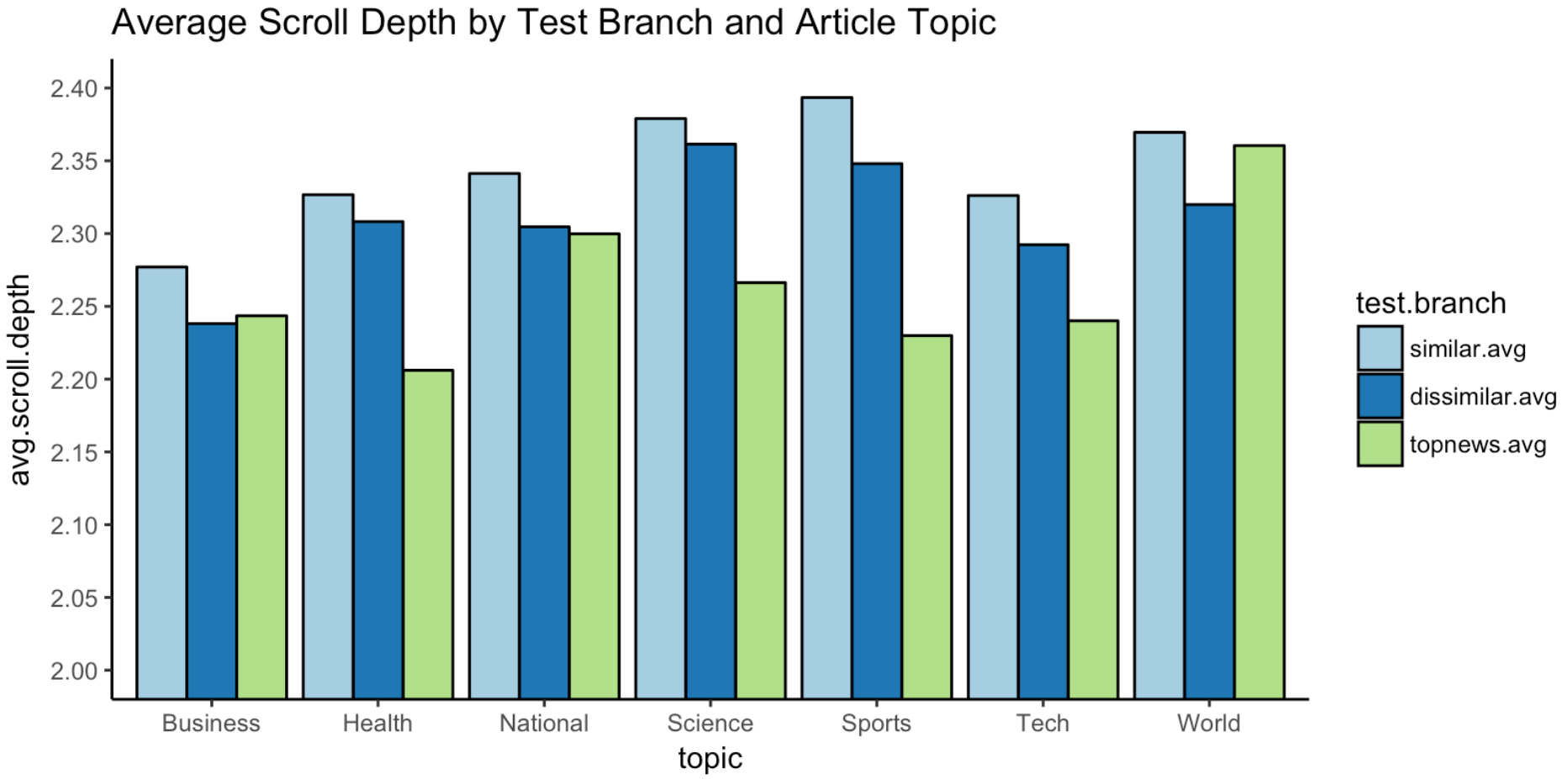
Source	Avg Scroll Depth	# of Winning Pages
similar	2.33	1,908
dissimilar	2.29	1,298
top news	2.29	1,351

- differences were consistent across all pages: similar scrolls were the “winners” against dissimilar and top news scrolls in **39% (1,908) of all article tests**

- top news scrolls won 28% (1,351) and dissimilar 27% (1,298); 6% were inconclusive

within topic differences

- trends held over all article topics, with greater differences in more niche areas, such as sports
- suggests that users who visit these more niche topics are inclined to read on and explore them in greater detail

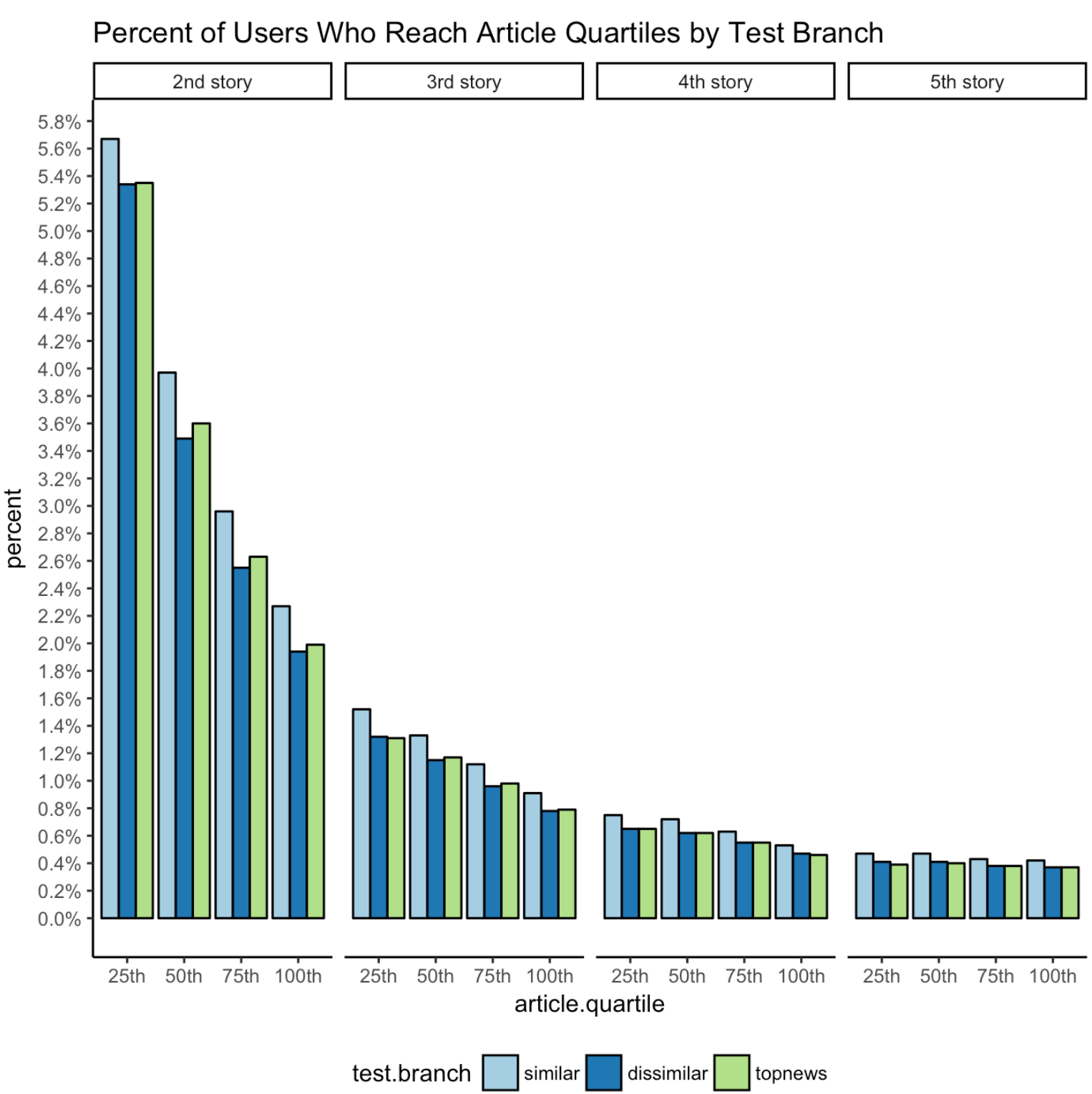


⬆️ article quartile depth

- article quartile depth: how deep users are getting into the articles that make up a scroll

- roughly 2.3% of users scrolled to the final quartile of the second article in similar scrolls, versus 1.9% and 2% for dissimilar and top news scrolls, respectively

- indicates users are also more engaged with the content when it is similarly related


Percent of Users Who Reach Article Quartiles by Test Branch
2nd story
3rd story
4th story
5th story
percent
5.8%
5.6%
5.4%
5.2%
5.0%
4.8%
4.6%
4.4%
4.2%
4.0%
3.8%
3.6%
3.4%
3.2%
3.0%
2.8%
2.6%
2.4%
2.2%
2.0%
1.8%
1.6%
1.4%
1.2%
1.0%
0.8%
0.6%
0.4%
0.2%
0.0%
25th
50th
75th
100th
article.quartile
test.branch
similar
dissimilar
topnews

the future (reuters version)

- personalization
- article length issues

the future (generally)

- embeddings 💀?
- “universal embeddings” for transfer learning
- ELMo (“Embeddings from Language Models”)
  - captures polysemy
  - character level training to handle OOV words

END 👋