Recommender System for Publisher of Technical News

Yawen Sun, Yue Li, Yuzi Luo
Mentor: Rok Sosic, Jeffrey Ullman

Our Problem

Digital Trends (https://www.digitaltrends.com)

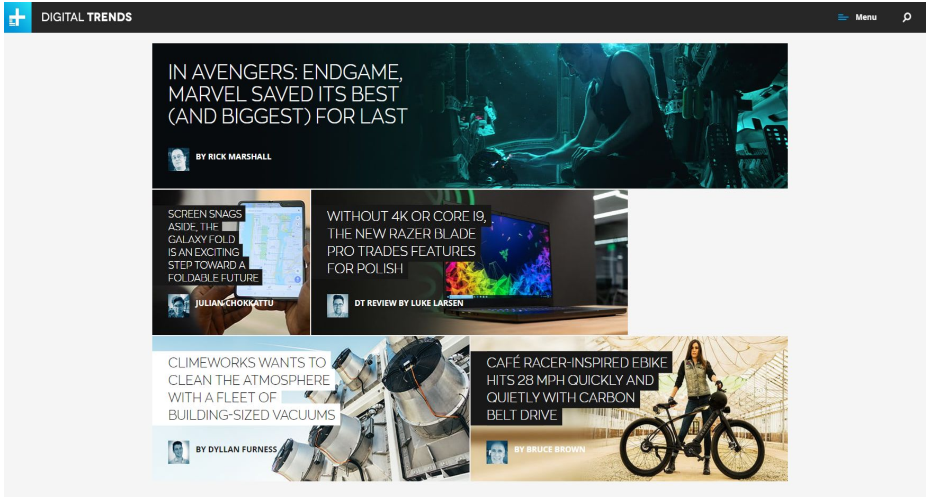
- Technology news, lifestyle, and information website.

Dataset

- Information about 170,000 News articles collected by content management system.
- Event data from real-time data collection platform over the past 15 months.

Goal

- Provide personalized content to users.



IN AVENGERS: ENDGAME, MARVEL SAVED ITS BEST (AND BIGGEST) FOR LAST

BY RICK MARSHALL

SCREEN SNAGS ASIDE, THE GALAXY FOLD IS AN EXCITING STEP TOWARD A FOLDABLE FUTURE

JULIAN CHOKKATTU

WITHOUT 4K OR CORE I9, THE NEW RAZER BLADE PRO TRADES FEATURES FOR POLISH

DT REVIEW BY LUKE LARSEN

CLIMEWORKS WANTS TO CLEAN THE ATMOSPHERE WITH A FLEET OF BUILDING-SIZED VACUUMS

BY DYLLAN FURNESS

CAFÉ RACER-INSPIRED EBIKE HITS 28 MPH QUICKLY AND QUIETLY WITH CARBON BELT DRIVE

BY BRUCE BROWN

Methodology - Feature Extraction

News Articles Feature Extraction

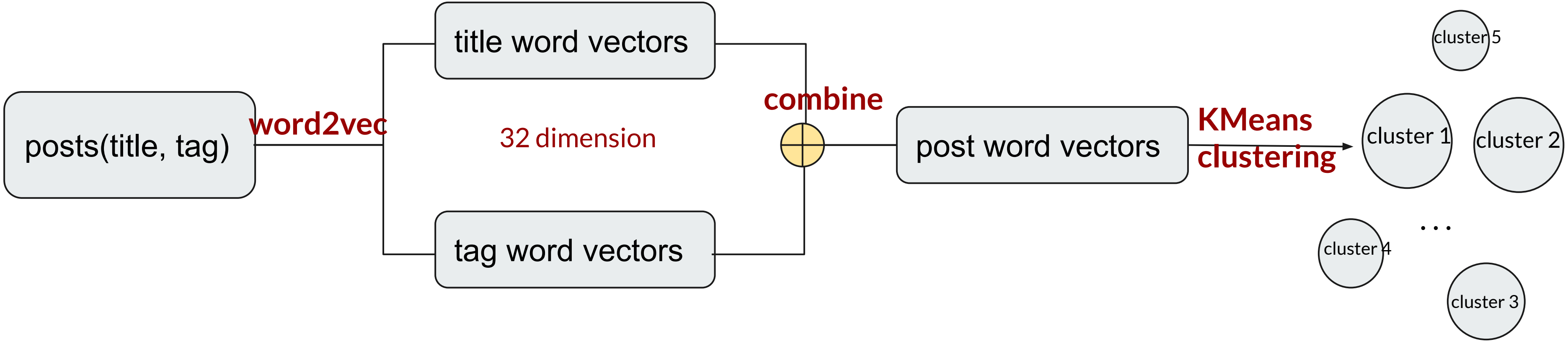
- Timestamp of Publish Date
- 32-D Vector Representing Keywords using Word2Vec
- Category using K-means Clustering

User Feature Extraction

- Location
- Number of Clicks from Each Cluster
- Timestamp of Each Click Event

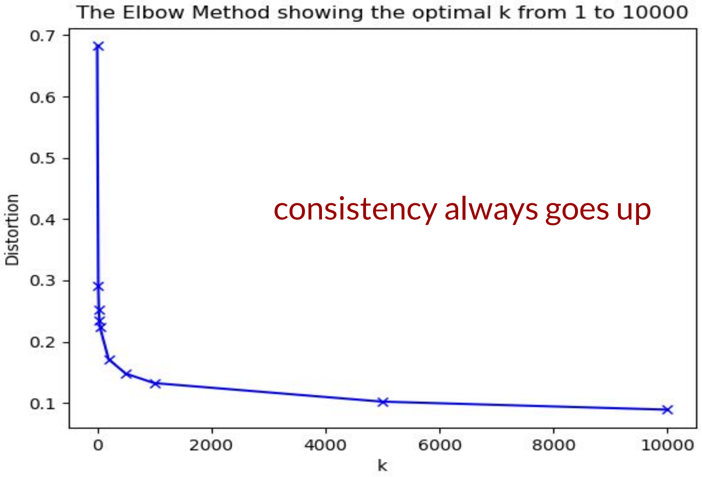
Posts > Word Vectors > Clusters

- clustering of posts by topic is our critical issue
- our clusters are generated based on tags and title words
- examination of clusters by either titles or tags show a lot of consistency

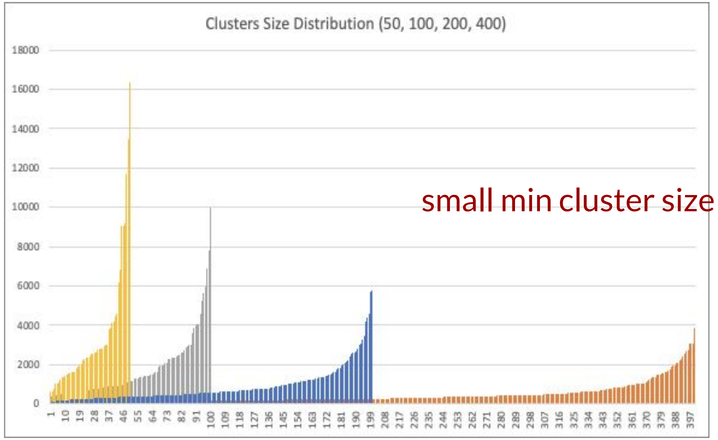


Clusters Matter

- consistency of clusters: **distortion ∝ 1/k**
- size of clusters: **size ∝ 1/k**



consistency always goes up



small min cluster size

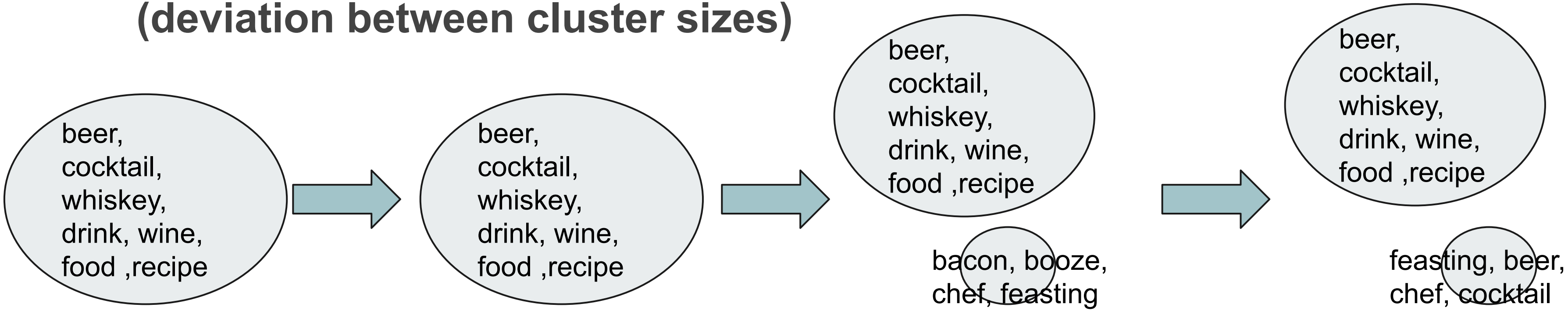
trade off between inter-cluster semantic consistency & cluster size distribution & computational resource

Clusters Optimization

- **used ‘elbow method’ to pick the optimal number of clusters (consistency vs. computation)**

- **customized stop words according to posts’ frequent but less meaningful words (everything you need to know about… the best…)**

- **re-clustered the above-average-size clusters (deviation between cluster sizes)**



beer, cocktail, whiskey, drink, wine, food ,recipe

beer, cocktail, whiskey, drink, wine, food ,recipe

beer, cocktail, whiskey, drink, wine, food ,recipe

bacon, booze, chef, feasting

beer, cocktail, whiskey, drink, wine, food ,recipe

feasting, beer, chef, cocktail

Experiment Setup

Training & Test Data

- Select 7,000 users frequently visiting the website over 15 months.
- First 9 months' viewing history for training
- last 6 months for val & test

Learning Model

- Similarity Learning $f_W(x, z) = x^T W z.$
- Wide and Deep combination of a linear model and a neural network

Evaluation

- Mean reciprocal rank (MRR)

  $$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}.$$

  - by highest rank of exact posts
  - by highest rank of similar (cosine similarity > 0.95) posts from the same cluster

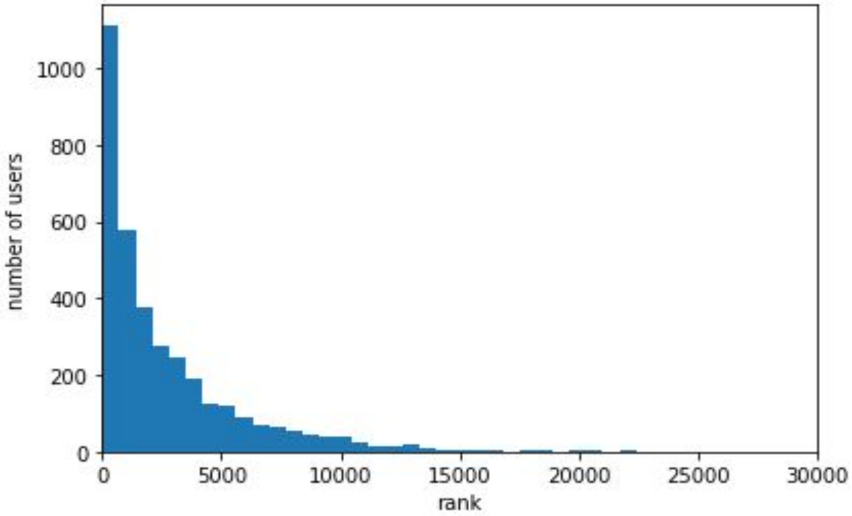
Results

cluster number = 50

exact posts:

SL:
val MRR: 0.0012
test MRR: 0.0012

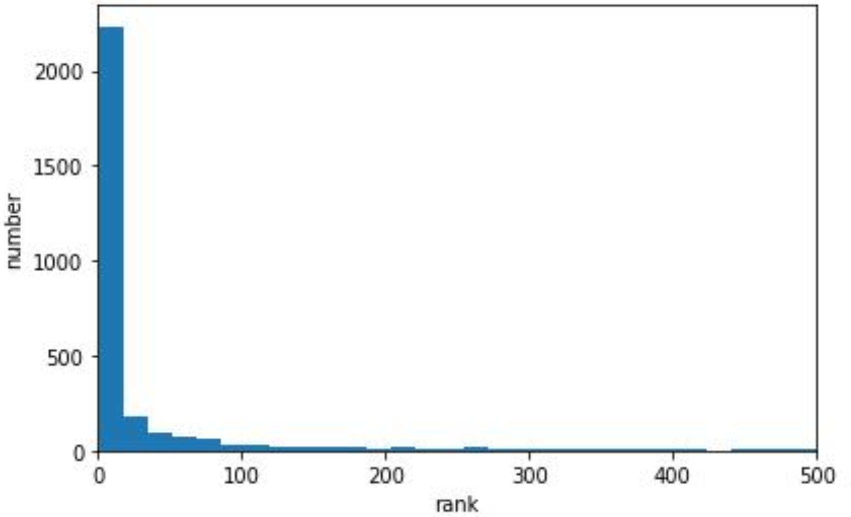
W&D:
val MRR: 0.023
test MRR: 0.021



similar posts:

SL:
val MRR: 0.0328
test MRR: 0.0356

W&D:
val MRR: 0.397
test MRR: 0.392



Results - evaluate by similarity

num of test users = 3500, threshold = 0.95

	num_cluster K	mrr	num_rank < 10	training time
similarity learning	50	0.0356	126	2h15min
	100	0.1111	1246	2h12min
	200	0.0743	1186	2h16min
	400	0.2068	1243	2h32min
wide and deep	50	0.3924	2112	7h03min
	100	0.4731	2261	8h38min
	200	0.4980	2225	12h02min
	400	0.2932	1356	19h37min

Limitations & Future Work

- **Lack of Article Content**
- **Lack of Negative Data Points**
- **More Factors in News Recommendation to Consider**
    - ❑ Short Time Big News Event
    - ❑ Cold Start Problem / Model Updating
    - ❑ Novelty Exploration