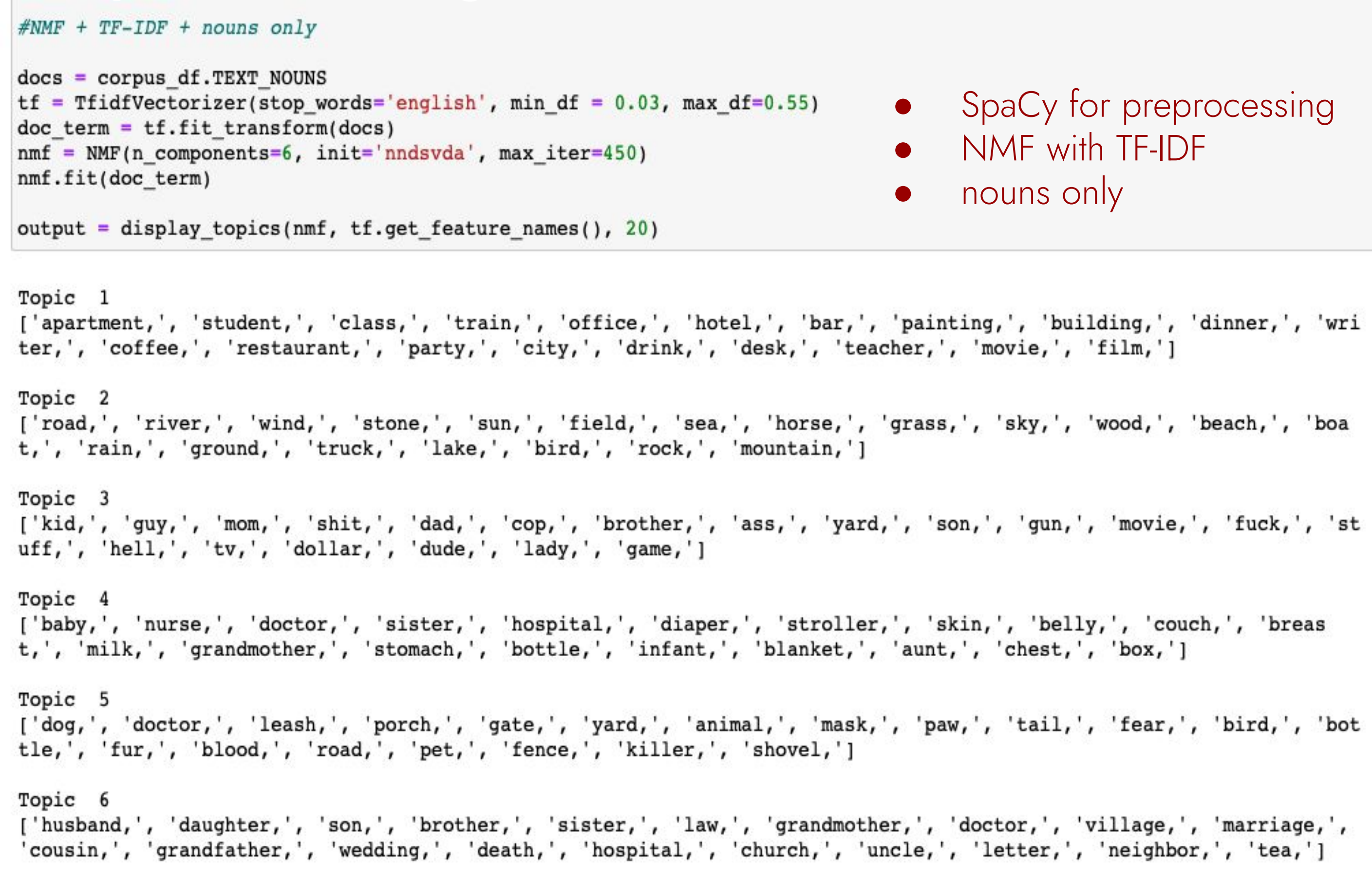
Short Story Recommender

Using NLP to analyze and recommend New Yorker short stories based on topic and writing style

Data

- **944** short stories from The New Yorker from **2002 to 2022**
- **333** different authors
- Top 5 authors:
  - Tessa Hadley (26 stories)
  - Thomas McGuane (20)
  - T.C. Boyle (19)
  - George Saunders (17)
  - Alice Munro (17)
- Word count range: **593** words to **16,122** words
- Average word count: **5,717** words

Topic Modeling



```
#NMF + TF-IDF + nouns only

docs = corpus_df.TEXT_NOUNS
tf = TfidfVectorizer(stop_words='english', min_df = 0.03, max_df=0.55)
doc_term = tf.fit_transform(docs)
nmf = NMF(n_components=6, init='nndsvda', max_iter=450)
nmf.fit(doc_term)

output = display_topics(nmf, tf.get_feature_names(), 20)
```

- SpaCy for preprocessing
- NMF with TF-IDF
- nouns only

```
Topic  1
['apartment,', 'student,', 'class,', 'train,', 'office,', 'hotel,', 'bar,', 'painting,', 'building,', 'dinner,', 'wri
ter,', 'coffee,', 'restaurant,', 'party,', 'city,', 'drink,', 'desk,', 'teacher,', 'movie,', 'film,']

Topic  2
['road,', 'river,', 'wind,', 'stone,', 'sun,', 'field,', 'sea,', 'horse,', 'grass,', 'sky,', 'wood,', 'beach,', 'boa
t,', 'rain,', 'ground,', 'truck,', 'lake,', 'bird,', 'rock,', 'mountain,']

Topic  3
['kid,', 'guy,', 'mom,', 'shit,', 'dad,', 'cop,', 'brother,', 'ass,', 'yard,', 'son,', 'gun,', 'movie,', 'fuck,', 'st
uff,', 'hell,', 'tv,', 'dollar,', 'dude,', 'lady,', 'game,']

Topic  4
['baby,', 'nurse,', 'doctor,', 'sister,', 'hospital,', 'diaper,', 'stroller,', 'skin,', 'belly,', 'couch,', 'breas
t,', 'milk,', 'grandmother,', 'stomach,', 'bottle,', 'infant,', 'blanket,', 'aunt,', 'chest,', 'box,']

Topic  5
['dog,', 'doctor,', 'leash,', 'porch,', 'gate,', 'yard,', 'animal,', 'mask,', 'paw,', 'tail,', 'fear,', 'bird,', 'bot
tle,', 'fur,', 'blood,', 'road,', 'pet,', 'fence,', 'killer,', 'shovel,']

Topic  6
['husband,', 'daughter,', 'son,', 'brother,', 'sister,', 'law,', 'grandmother,', 'doctor,', 'village,', 'marriage,',
'cousin,', 'grandfather,', 'wedding,', 'death,', 'hospital,', 'church,', 'uncle,', 'letter,', 'neighbor,', 'tea,']
```

Topic 1: City/School

Topic 2: Country/Nature?

Topic 3: Dude???

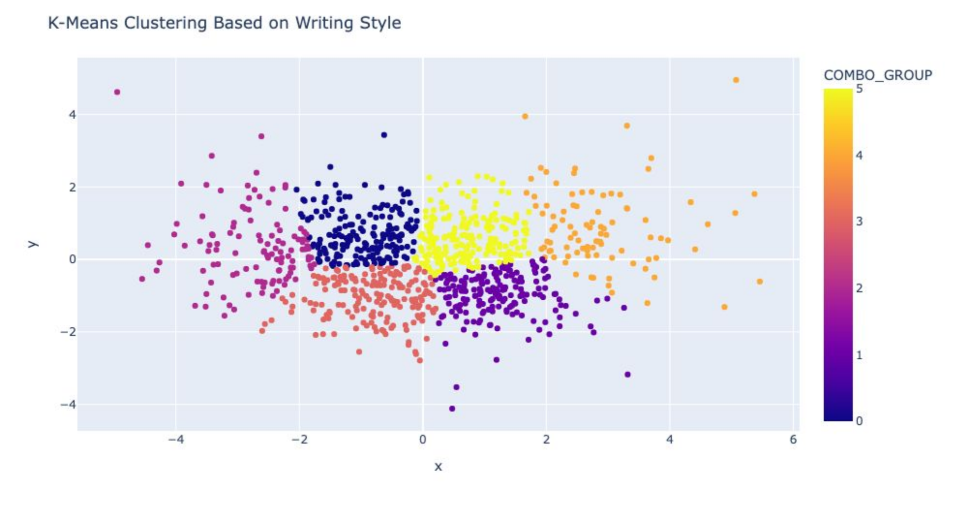
Topic 4: Pregnancy?

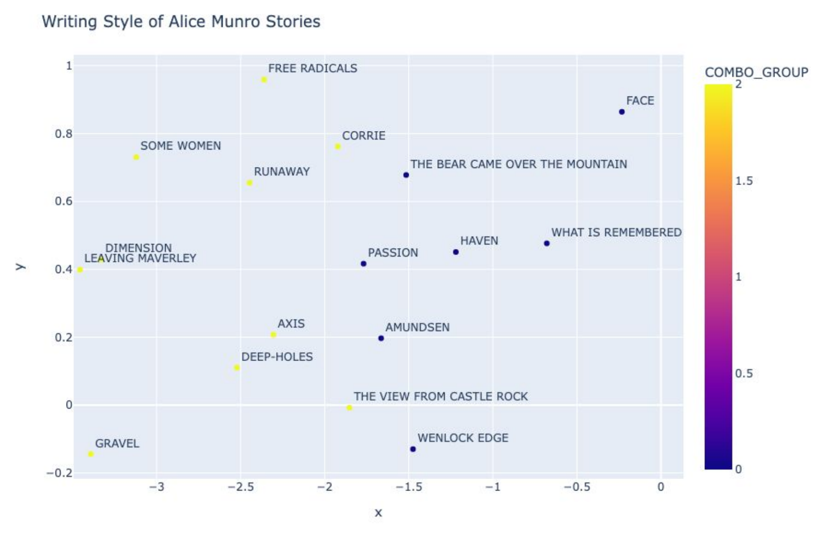
Topic 5: Animals

Topic 6: Family

Clustering Around Writing Style

- Used SpaCy POS tagging to count nouns, verbs, adjs, advs
- Used LexicalRichness module to find Type-Token Ratio (TTR)
  - unique words / total words
- Calculated average # of words per sentence
- Reduced dimensions with PCA
- K-means clustering with 6 clusters
- Assigned each story to a cluster (“combo group”)





Short Story Recommender

User chooses:

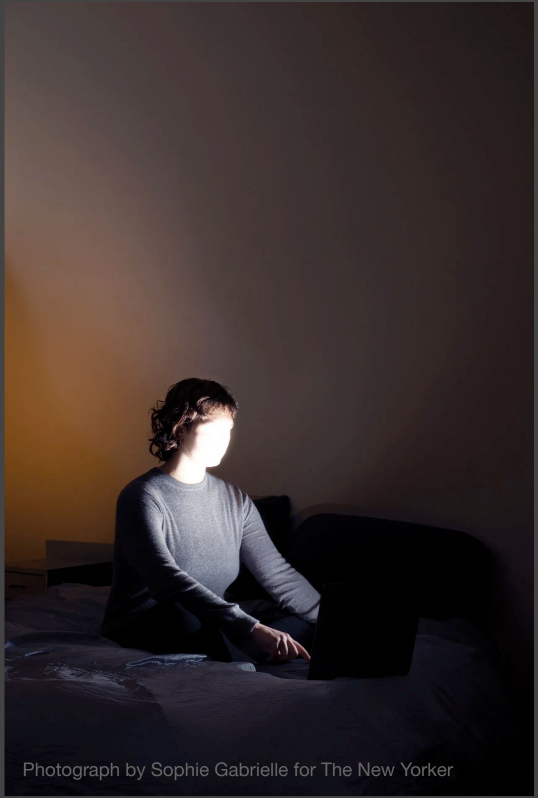
- Short story index #
- Number of recommendations to get
- Weight of topic similarity
- Weight of style match
- Weight of author match

Recommendation Score =

topic similarity score (cosine similarity)

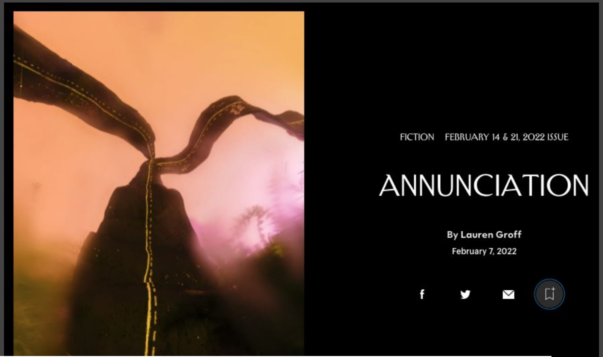
+ author match weight x author match (1 or 0)

+ style match weight x style match (1 or 0)



Photograph by Sophie Gabrielle for The New Yorker

Example: Annunciation by Lauren Groff



FICTION FEBRUARY 14 & 21, 2022 ISSUE

ANNUNCIATION

By Lauren Groff

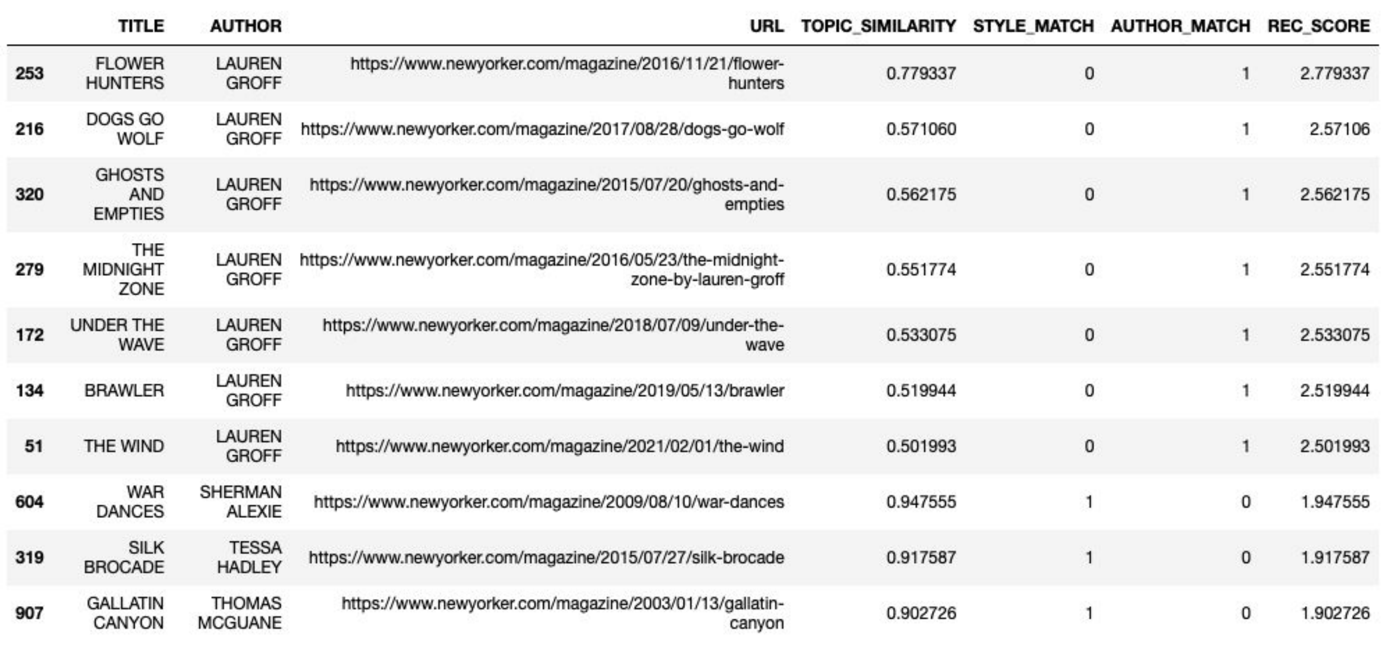
February 7, 2022

# of results: 10

Topic weight: 1

Style weight: 1

Author weight: 2



	TITLE	AUTHOR	URL	TOPIC_SIMILARITY	STYLE_MATCH	AUTHOR_MATCH	REC_SCORE
253	FLOWER HUNTERS	LAUREN GROFF	https://www.newyorker.com/magazine/2016/11/21/flower-hunters	0.779337	0	1	2.779337
216	DOGS GO WOLF	LAUREN GROFF	https://www.newyorker.com/magazine/2017/08/28/dogs-go-wolf	0.571060	0	1	2.57106
320	GHOSTS AND EMPTIES	LAUREN GROFF	https://www.newyorker.com/magazine/2015/07/20/ghosts-and-empties	0.562175	0	1	2.562175
279	THE MIDNIGHT ZONE	LAUREN GROFF	https://www.newyorker.com/magazine/2016/05/23/the-midnight-zone-by-lauren-groff	0.551774	0	1	2.551774
172	UNDER THE WAVE	LAUREN GROFF	https://www.newyorker.com/magazine/2018/07/09/under-the-wave	0.533075	0	1	2.533075
134	BRAWLER	LAUREN GROFF	https://www.newyorker.com/magazine/2019/05/13/brawler	0.519944	0	1	2.519944
51	THE WIND	LAUREN GROFF	https://www.newyorker.com/magazine/2021/02/01/the-wind	0.501993	0	1	2.501993
604	WAR DANCES	SHERMAN ALEXIE	https://www.newyorker.com/magazine/2009/08/10/war-dances	0.947555	1	0	1.947555
319	SILK BROCADE	TESSA HADLEY	https://www.newyorker.com/magazine/2015/07/27/silk-brocade	0.917587	1	0	1.917587
907	GALLATIN CANYON	THOMAS MCGUANE	https://www.newyorker.com/magazine/2003/01/13/gallatin-canyon	0.902726	1	0	1.902726

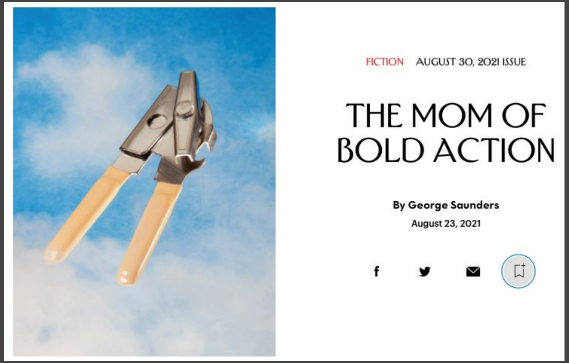
Example: The Mom of Bold Action by George Saunders

# of results: 5

Topic weight: 1

Style weight: 0.5

Author weight: 0

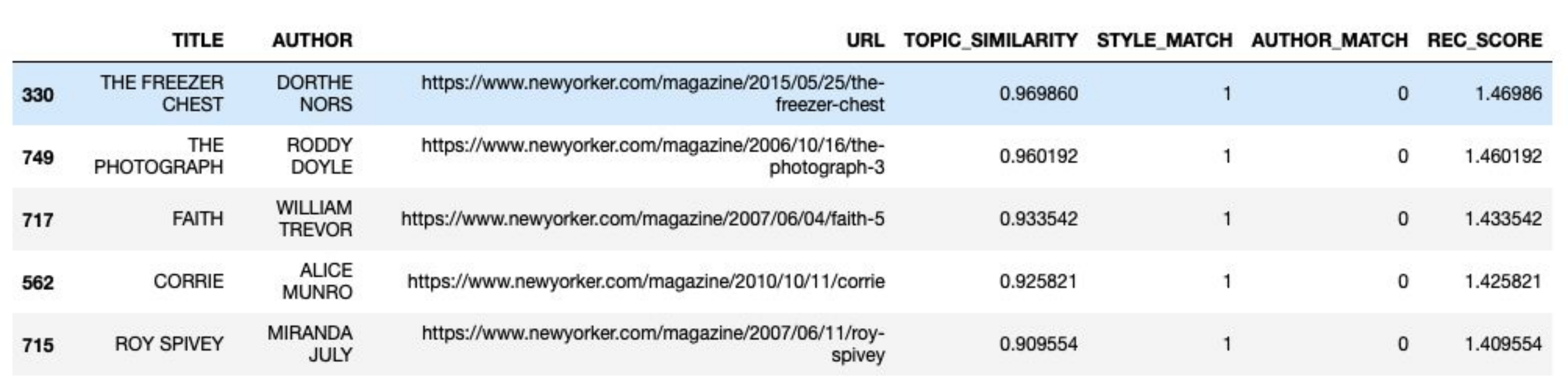


FICTION AUGUST 30, 2021 ISSUE

THE MOM OF BOLD ACTION

By George Saunders

August 23, 2021



	TITLE	AUTHOR	URL	TOPIC_SIMILARITY	STYLE_MATCH	AUTHOR_MATCH	REC_SCORE
330	THE FREEZER CHEST	DORTHE NORS	https://www.newyorker.com/magazine/2015/05/25/the-freezer-chest	0.969860	1	0	1.46986
749	THE PHOTOGRAPH	RODDY DOYLE	https://www.newyorker.com/magazine/2006/10/16/the-photograph-3	0.960192	1	0	1.460192
717	FAITH	WILLIAM TREVOR	https://www.newyorker.com/magazine/2007/06/04/faith-5	0.933542	1	0	1.433542
562	CORRIE	ALICE MUNRO	https://www.newyorker.com/magazine/2010/10/11/corrie	0.925821	1	0	1.425821
715	ROY SPIVEY	MIRANDA JULY	https://www.newyorker.com/magazine/2007/06/11/roy-spivey	0.909554	1	0	1.409554

Future Steps

- Improve method of calculating recommendation score
- Refine matching based on writing style—search for closest data points instead? Figure out different ways of measuring writing style?
- Simplify user inputs to make them more intuitive
- Build a Streamlit app