

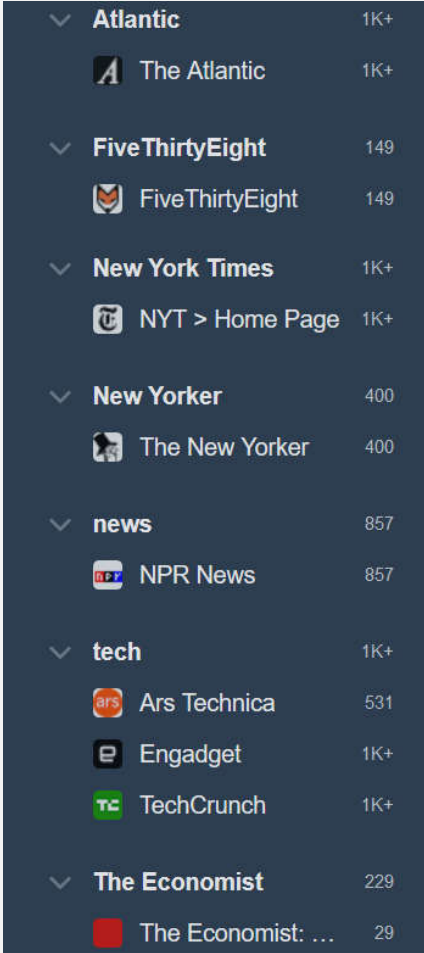











HEART-ICLE

JAMES HU

MOTIVATION

- Too much to read
- Hundreds of articles
- Dozens of news channels
- RSS is going out of control



▼ Atlantic	1K+
 The Atlantic	1K+
▼ FiveThirtyEight	149
 FiveThirtyEight	149
▼ New York Times	1K+
 NYT > Home Page	1K+
▼ New Yorker	400
 The New Yorker	400
▼ news	857
 NPR News	857
▼ tech	1K+
 Ars Technica	531
 Engadget	1K+
 TechCrunch	1K+
▼ The Economist	229
 The Economist: ...	29

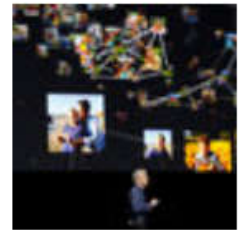
SUMMARIES SUCK

- Clearly designed for clicks rather than give you the story
- News is looking for ad revenue

With Sales Sluggish, Apple Revamps Its Software

By VINDU GOEL 6:48 PM ET

In an effort to keep developers happily writing apps for its software, as well as to reinvigorate its products, Apple will allow access to two of its crown jewels, hoping to spur creativity.



• **Highlights and Analysis From Apple's Conference** 3:31 PM ET

MEDIATOR

Preparing to Stand Up to Sports Incorporated

By JIM RUTENBERG

Bill Simmons, the influential commentator who was fired from ESPN after making incendiary remarks concerning the National Football League, begins a new show on HBO on June 22.



CONCEPT

- Most summarization algorithms are unsupervised and subpar
- Medium is a popular blogging platform where users can select highlights from the article
- Highlights are aggregated and the best is shown

MEDIUM HIGHLIGHT EXAMPLE

Being tired isn't a badge of honor

Whenever I speak at a conference, I try to catch a few of the other presentations. I tend to stand in the back and listen, observe, and get a general sense of the room.

Lately, I've been hearing something that disturbs me. A lot of entrepreneurs onstage have been bragging about not sleeping, telling their audiences about their 16-hour days, and making it sound like hustle-at-all-costs is the way ahead. Rest be damned, they say—there's an endless amount of work to do. *

I think this message is one of the most harmful in all of business. Sustained exhaustion is not a rite of passage. It's a mark of stupidity. Literally. Scientists have suggested that scores on IQ tests decline on each successive day you sleep less than you naturally would. It doesn't take long before the difference is telling. * Top highlight

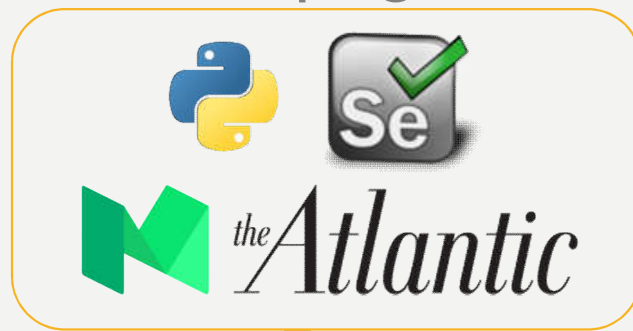
Crowdsourced
Highlight

THE GOAL

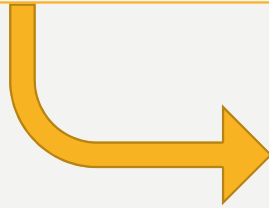
- Medium highlights represent what people consider the most important part or “heart” of an article
- Collect this information and use Doc2Vec to capture the meaning of what a highlight is compared to its document
- Use these highlights to find the “heart” in other articles

PIPELINE

Scraping



Modeling

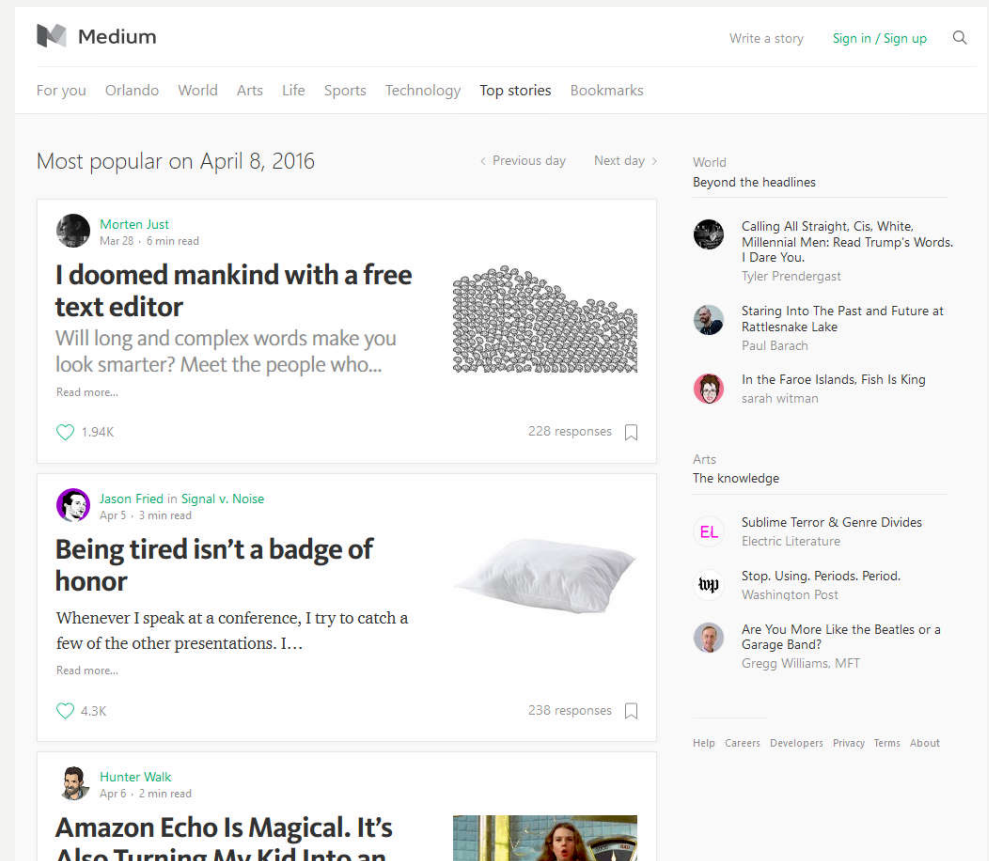




SCRAPING

SCRAPING MEDIUM

- Scraped most popular daily pages since September 2014 for popular posts
- Acquired raw list of 4000+ documents

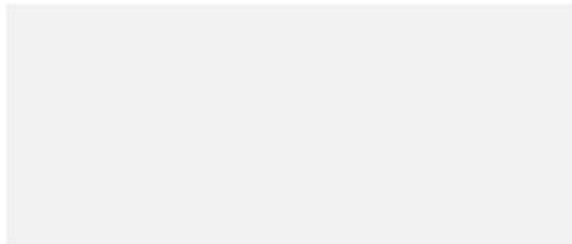


The screenshot displays the Medium website interface. At the top, the Medium logo is on the left, and navigation links for "Write a story", "Sign in / Sign up", and a search icon are on the right. Below the header, there are category tabs: "For you", "Orlando", "World", "Arts", "Life", "Sports", "Technology", "Top stories", and "Bookmarks". The main content area is titled "Most popular on April 8, 2016" and includes navigation for "Previous day" and "Next day". Three featured posts are visible:

- Post 1:** By Morten Just (Mar 28 · 6 min read). Title: "I doomed mankind with a free text editor". Subtitle: "Will long and complex words make you look smarter? Meet the people who...". Engagement: 1.94K likes, 228 responses. Image: A large pile of coins.
- Post 2:** By Jason Fried (Apr 5 · 3 min read). Title: "Being tired isn't a badge of honor". Subtitle: "Whenever I speak at a conference, I try to catch a few of the other presentations. I...". Engagement: 4.3K likes, 238 responses. Image: A white pillow.
- Post 3:** By Hunter Walk (Apr 6 · 2 min read). Title: "Amazon Echo Is Magical. It's Also Turning My Kid Into an...". Image: A young girl looking at a screen.

On the right side, there are vertical sections for "World: Beyond the headlines" (listing articles like "Calling All Straight, Cis, White, Millennial Men: Read Trump's Words. I Dare You." by Tyler Prendergast), "Arts: The knowledge" (listing "Sublime Terror & Genre Divides" by Electric Literature and "Stop. Using. Periods. Period." by Washington Post), and a footer with links for "Help", "Careers", "Developers", "Privacy", "Terms", and "About".

SCRAPING MEDIUM



Being tired isn't a badge of honor

Whenever I speak at a conference, I try to catch a few of the other presentations. I tend to stand in the back and listen, observe, and get a general sense of the room.

Lately, I've been hearing something that disturbs me. A lot of entrepreneurs onstage have been bragging about not sleeping, telling their audiences about their 16-hour days, and making it sound like hustle-at-all-costs is the way ahead. Rest be damned, they say—there's an endless amount of work to do.

I think this message is one of the most harmful in all of business. Sustained exhaustion is not a rite of passage. It's a mark of stupidity. Literally. Scientists have suggested that scores on IQ tests decline on each successive day you sleep less than you naturally would. It doesn't take long before the difference



Being tired isn't a badge of honor

Whenever I speak at a conference, I try to catch a few of the other presentations. I tend to stand in the back and listen, observe, and get a general sense of the room.

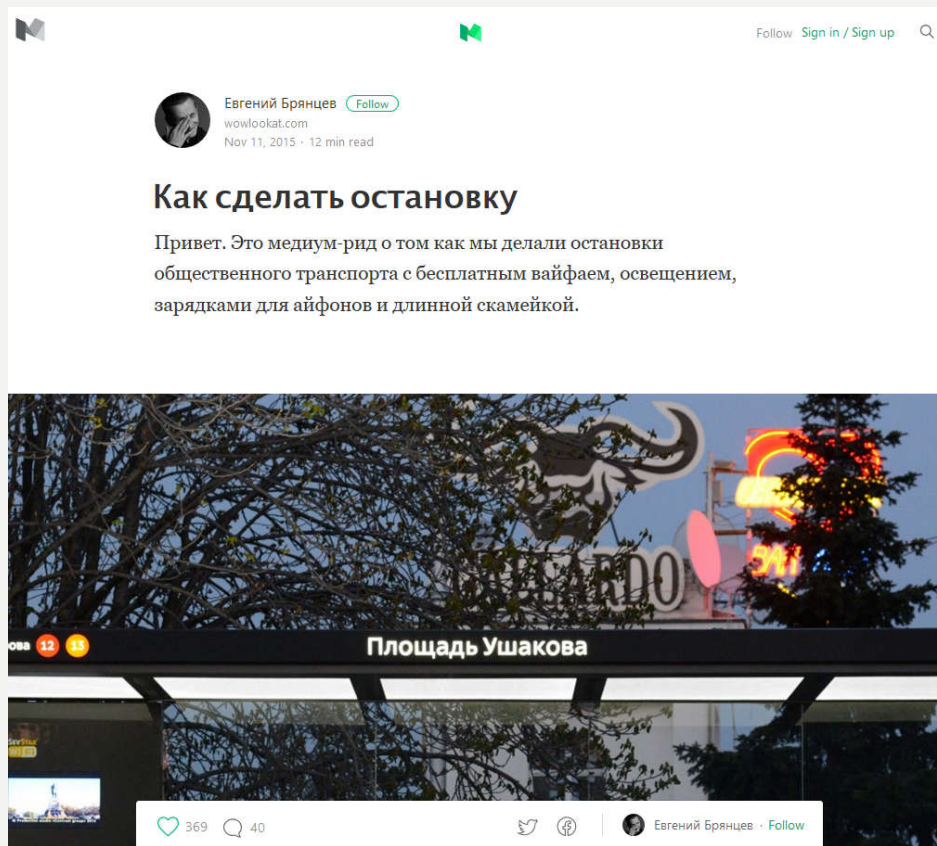
Lately, I've been hearing something that disturbs me. A lot of entrepreneurs onstage have been bragging about not sleeping, telling their audiences about their 16-hour days, and making it sound like hustle-at-all-costs is the way ahead. Rest be damned, they say—there's an endless amount of work to do.

I think this message is one of the most harmful in all of business. Sustained exhaustion is not a rite of passage. It's a mark of stupidity. Literally. Scientists

4.3K 238 Jason Fried · Follow

- Lots of Javascript loading required the use of Selenium
- Delay to load Javascript then save source

SCRAPING MEDIUM

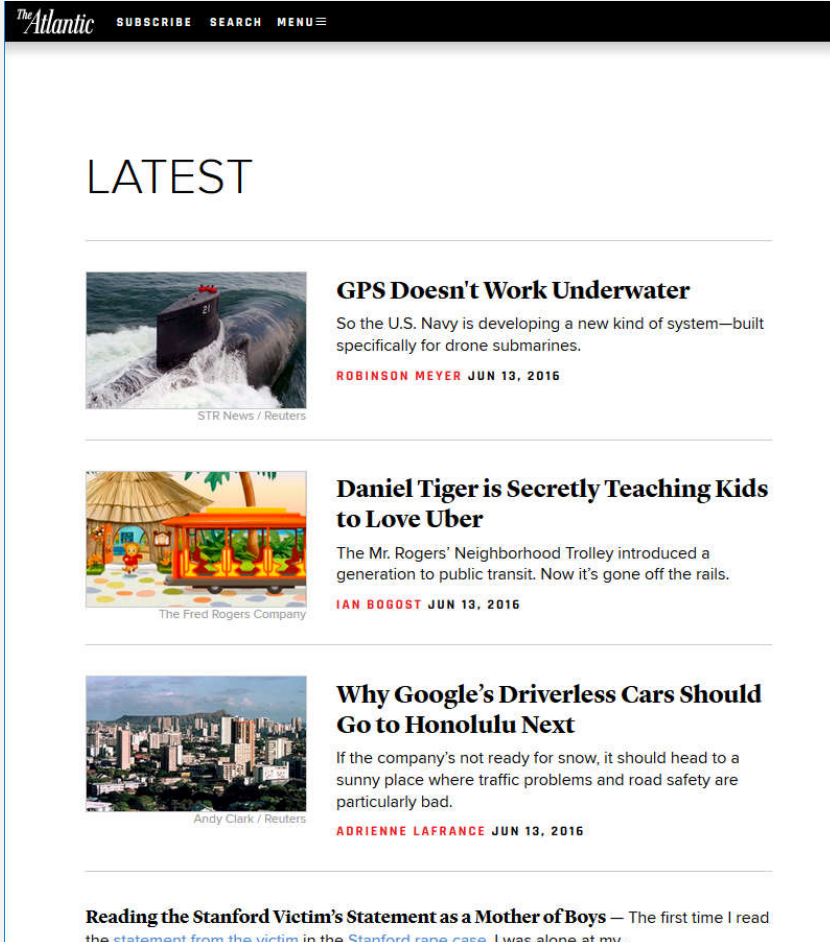


The screenshot shows a Medium article interface. At the top, there are navigation links for 'Follow', 'Sign in / Sign up', and a search icon. The author's profile is visible, including a circular profile picture, the name 'Евгений Бряцев', a 'Follow' button, the website 'wowlookat.com', and the date 'Nov 11, 2015 · 12 min read'. The article title is 'Как сделать остановку'. The main text reads: 'Привет. Это медиум-рид о том как мы делали остановки общественного транспорта с бесплатным вайфаем, освещением, зарядками для айфонов и длинной скамейкой.' Below the text is a photograph of a bus stop at night, with a sign that says 'Площадь Ушакова'. The photo shows trees and a brightly lit sign with a stylized bird logo. At the bottom of the article, there are engagement metrics: a heart icon with '369', a comment icon with '40', and social sharing icons for Twitter and Facebook. The author's name and 'Follow' button are repeated at the bottom right of the article content.

- Lots of bad data
- Foreign languages (especially Russian)
- Extremely short highlights
- Extremely short articles
- Parsed using BeautifulSoup


SCRAPING THE ATLANTIC


- Not enough articles in Medium corpus
- Used The Atlantic as an easy to get source of words and articles
- Easy to scrape, done with Requests and BeautifulSoup




The Atlantic SUBSCRIBE SEARCH MENU

LATEST

**GPS Doesn't Work Underwater**
So the U.S. Navy is developing a new kind of system—built specifically for drone submarines.
ROBINSON MEYER JUN 13, 2016
STR News / Reuters

**Daniel Tiger is Secretly Teaching Kids to Love Uber**
The Mr. Rogers' Neighborhood Trolley introduced a generation to public transit. Now it's gone off the rails.
IAN BOGOST JUN 13, 2016
The Fred Rogers Company

**Why Google's Driverless Cars Should Go to Honolulu Next**
If the company's not ready for snow, it should head to a sunny place where traffic problems and road safety are particularly bad.
ADRIENNE LAFRANCE JUN 13, 2016
Andy Clark / Reuters

Reading the Stanford Victim's Statement as a Mother of Boys — The first time I read the [statement from the victim](#) in the [Stanford rape case](#). I was alone at mv..



MODELING

GENSIM

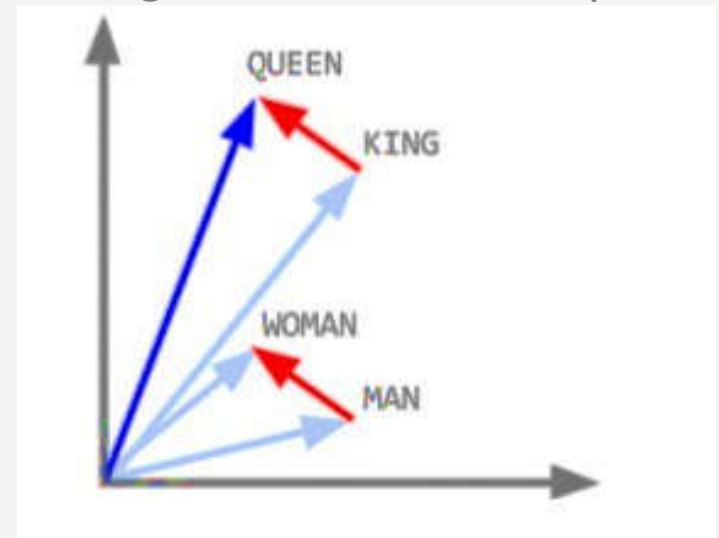


- One of the most powerful python packages for Natural Language Processing
- Selected for robust implementation of Doc2Vec

WORD2VEC

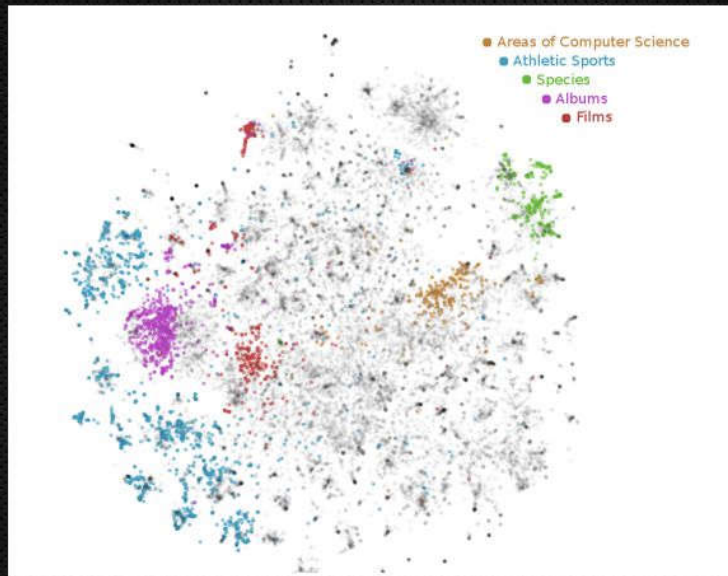
- Idea that words can be represented as vectors
- Words have vector relationships compared to each other

So king - man + woman = queen!



DOC2VEC

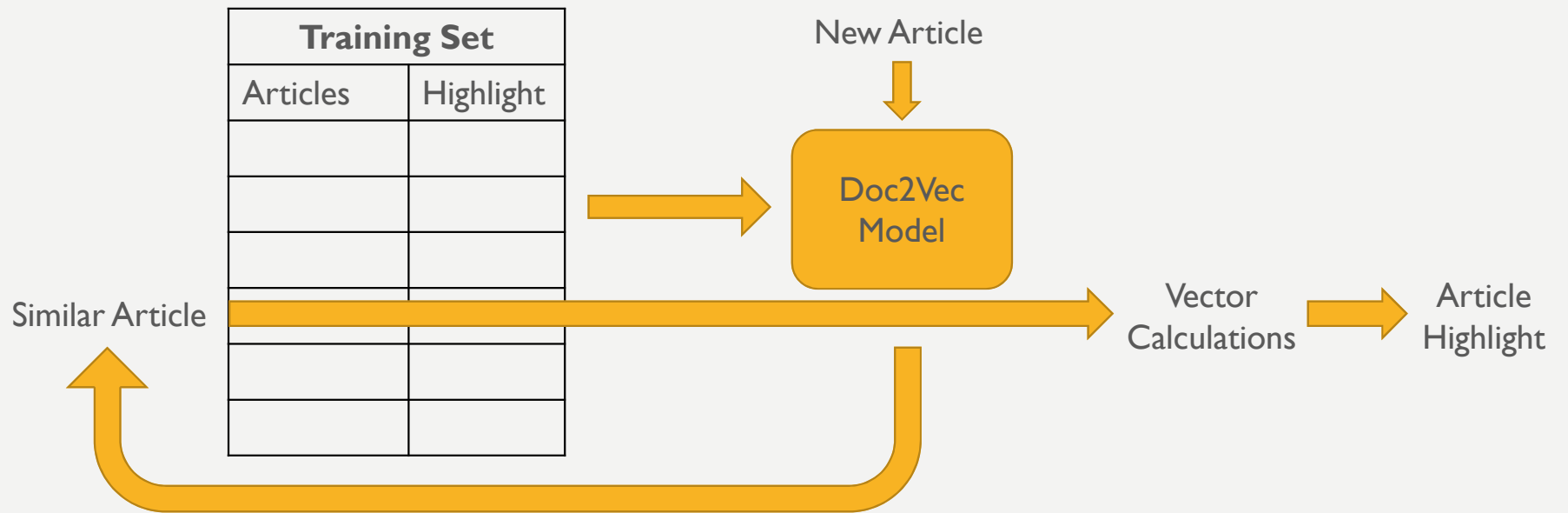
Doc2Vec on Wikipedia⁸



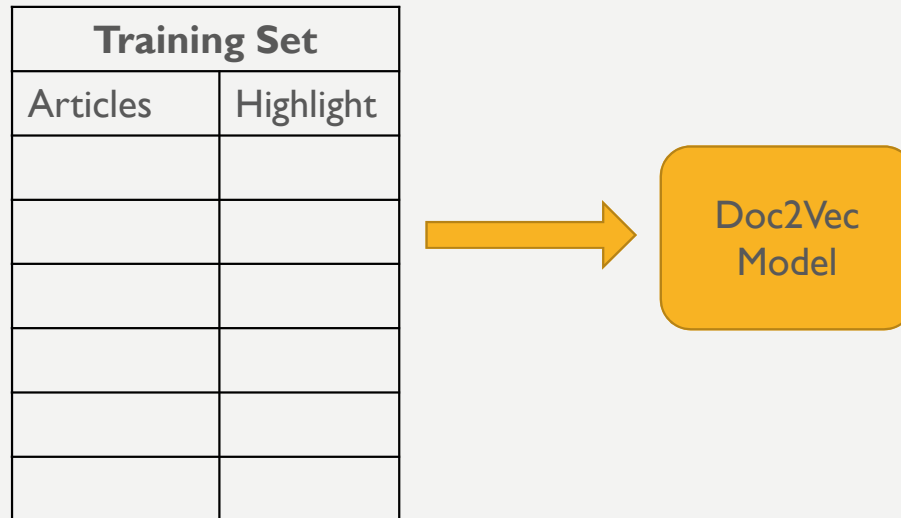
⁸Document Embedding with Paragraph Vectors, A. Dai et al., 2014

- Extension of Word2Vec
- Combine full document and its context to create vector representations

APPLYING DOC2VEC



TRAINING DOC2VEC



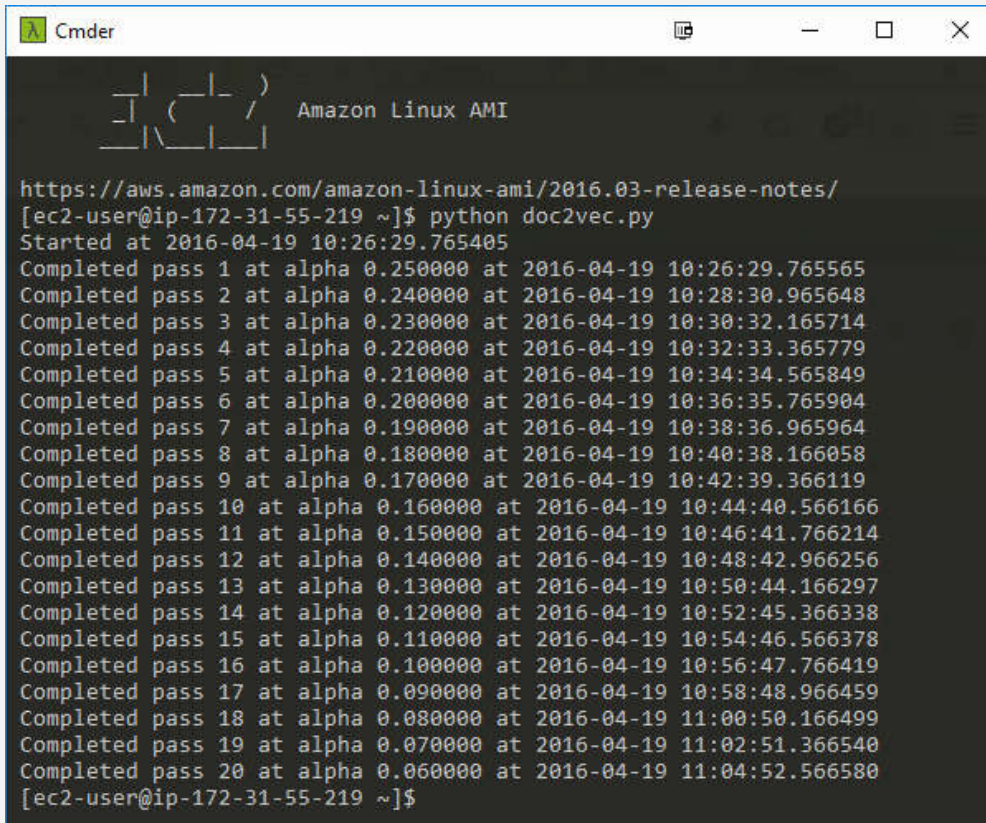
- Train Doc2Vec model using Medium and The Atlantic articles
- Requires significant cleaning to get data in right format
- Long run-times required use of AWS

CLEANING THE DATA

- Had to detect problems as indicated before
- Needed to reformat punctuation

```
43 # Get paragraphs
44 paragraphs = article.find_all('p')
45 paragraphs = map(Lambda x: unicode.unidecode(x.get_text().strip()), paragraphs)
46
47 paragraphsPrep = paragraphs
48 paragraphsPrep = map(Lambda x: x.lower(), paragraphsPrep)
49 paragraphsPrep = map(Lambda x: re.sub('\.', ' ', x), paragraphsPrep)
50 paragraphsPrep = map(Lambda x: re.sub('\\"', ' " ', x), paragraphsPrep)
51 paragraphsPrep = map(Lambda x: re.sub('\,', ' , ', x), paragraphsPrep)
52 paragraphsPrep = map(Lambda x: re.sub('\(', ' ( ', x), paragraphsPrep)
53 paragraphsPrep = map(Lambda x: re.sub('\)', ' ) ', x), paragraphsPrep)
54 paragraphsPrep = map(Lambda x: re.sub('\!', ' ! ', x), paragraphsPrep)
55 paragraphsPrep = map(Lambda x: re.sub('\?', ' ? ', x), paragraphsPrep)
56 paragraphsPrep = map(Lambda x: re.sub('\:', ' : ', x), paragraphsPrep)
57 paragraphsPrep = map(Lambda x: re.sub('\;', ' ; ', x), paragraphsPrep)
58 paragraphsPrep = map(Lambda x: re.sub('\-\-', ' -- ', x), paragraphsPrep)
59 paragraphsPrep = map(Lambda x: x.strip(), paragraphsPrep)
60
61 # Get text information, clean up unicode
62 articleClean = unicode.unidecode(" ".join(item.strip() for item in article.find_all(text=True)))
63
64 articleClean = articleClean.lower()
65 articleClean = re.sub('\.', ' ', articleClean)
66 articleClean = re.sub('\\"', ' " ', articleClean)
67 articleClean = re.sub('\,', ' , ', articleClean)
68 articleClean = re.sub('\(', ' ( ', articleClean)
```

RUNNING USING AWS

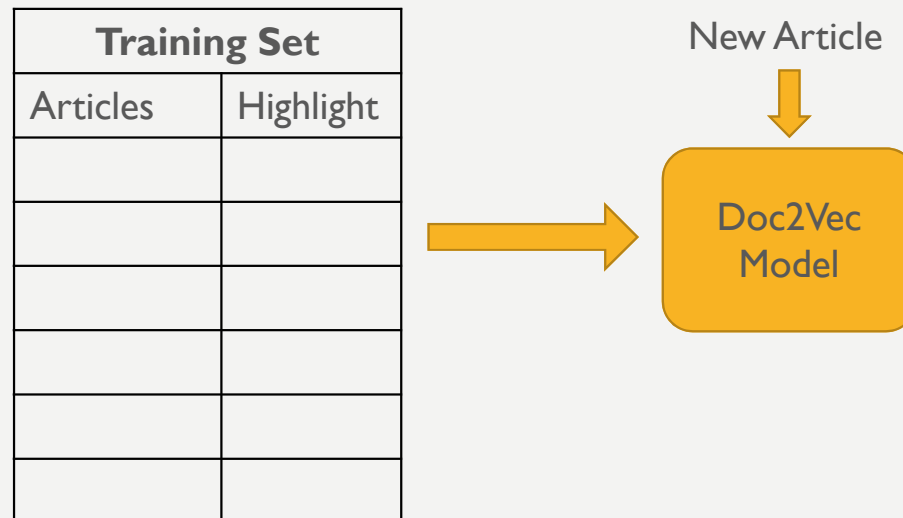


```
Cmdr
Amazon Linux AMI

https://aws.amazon.com/amazon-linux-ami/2016.03-release-notes/
[ec2-user@ip-172-31-55-219 ~]$ python doc2vec.py
Started at 2016-04-19 10:26:29.765405
Completed pass 1 at alpha 0.250000 at 2016-04-19 10:26:29.765565
Completed pass 2 at alpha 0.240000 at 2016-04-19 10:28:30.965648
Completed pass 3 at alpha 0.230000 at 2016-04-19 10:30:32.165714
Completed pass 4 at alpha 0.220000 at 2016-04-19 10:32:33.365779
Completed pass 5 at alpha 0.210000 at 2016-04-19 10:34:34.565849
Completed pass 6 at alpha 0.200000 at 2016-04-19 10:36:35.765904
Completed pass 7 at alpha 0.190000 at 2016-04-19 10:38:36.965964
Completed pass 8 at alpha 0.180000 at 2016-04-19 10:40:38.166058
Completed pass 9 at alpha 0.170000 at 2016-04-19 10:42:39.366119
Completed pass 10 at alpha 0.160000 at 2016-04-19 10:44:40.566166
Completed pass 11 at alpha 0.150000 at 2016-04-19 10:46:41.766214
Completed pass 12 at alpha 0.140000 at 2016-04-19 10:48:42.966256
Completed pass 13 at alpha 0.130000 at 2016-04-19 10:50:44.166297
Completed pass 14 at alpha 0.120000 at 2016-04-19 10:52:45.366338
Completed pass 15 at alpha 0.110000 at 2016-04-19 10:54:46.566378
Completed pass 16 at alpha 0.100000 at 2016-04-19 10:56:47.766419
Completed pass 17 at alpha 0.090000 at 2016-04-19 10:58:48.966459
Completed pass 18 at alpha 0.080000 at 2016-04-19 11:00:50.166499
Completed pass 19 at alpha 0.070000 at 2016-04-19 11:02:51.366540
Completed pass 20 at alpha 0.060000 at 2016-04-19 11:04:52.566580
[ec2-user@ip-172-31-55-219 ~]$
```

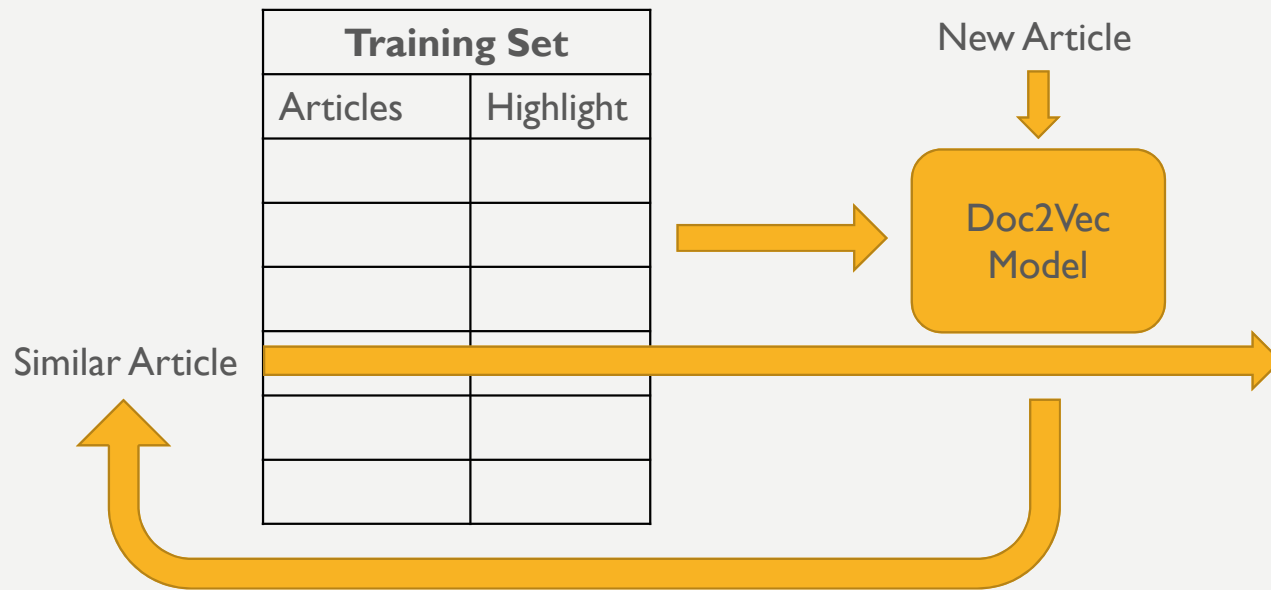
- Training models took several hours on local machine
- EC2 Instance c3.8xlarge with 32 cores completed training in ~40 minutes

INPUTTING NEW ARTICLES



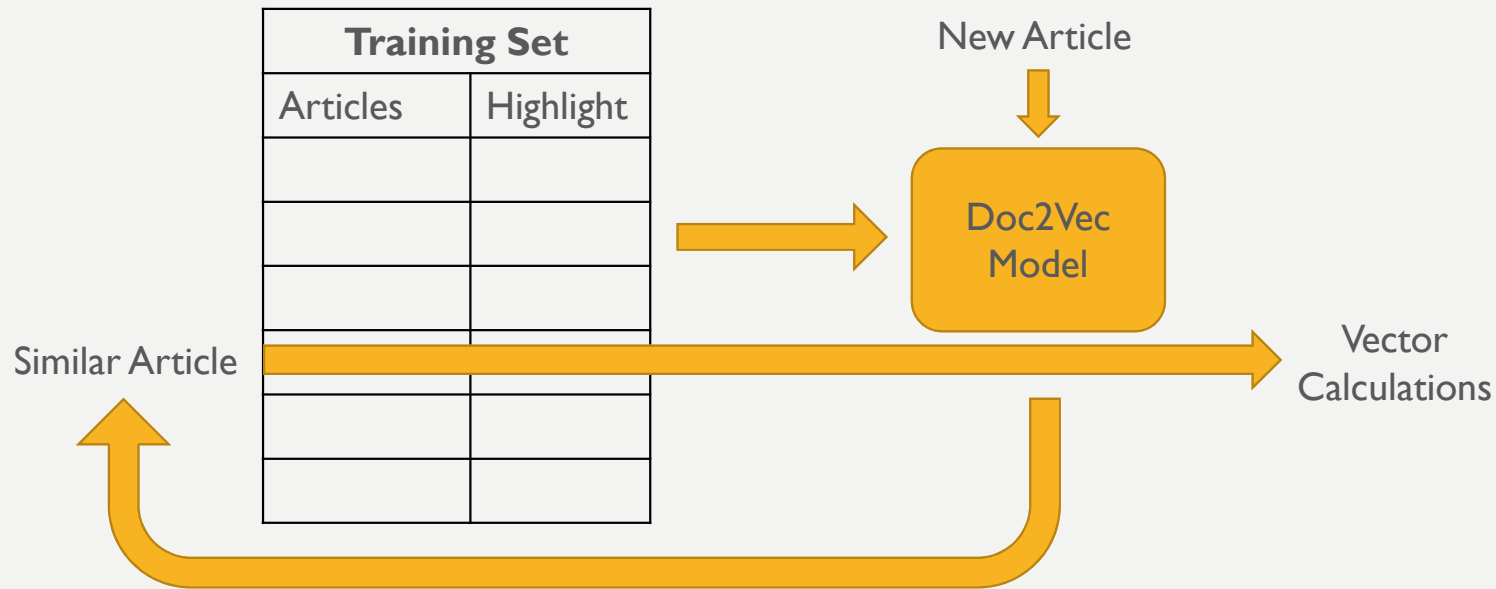
- A new article, not in the original corpus is inputted into the Doc2Vec model, outputting a vector

FINDING SIMILAR ARTICLES



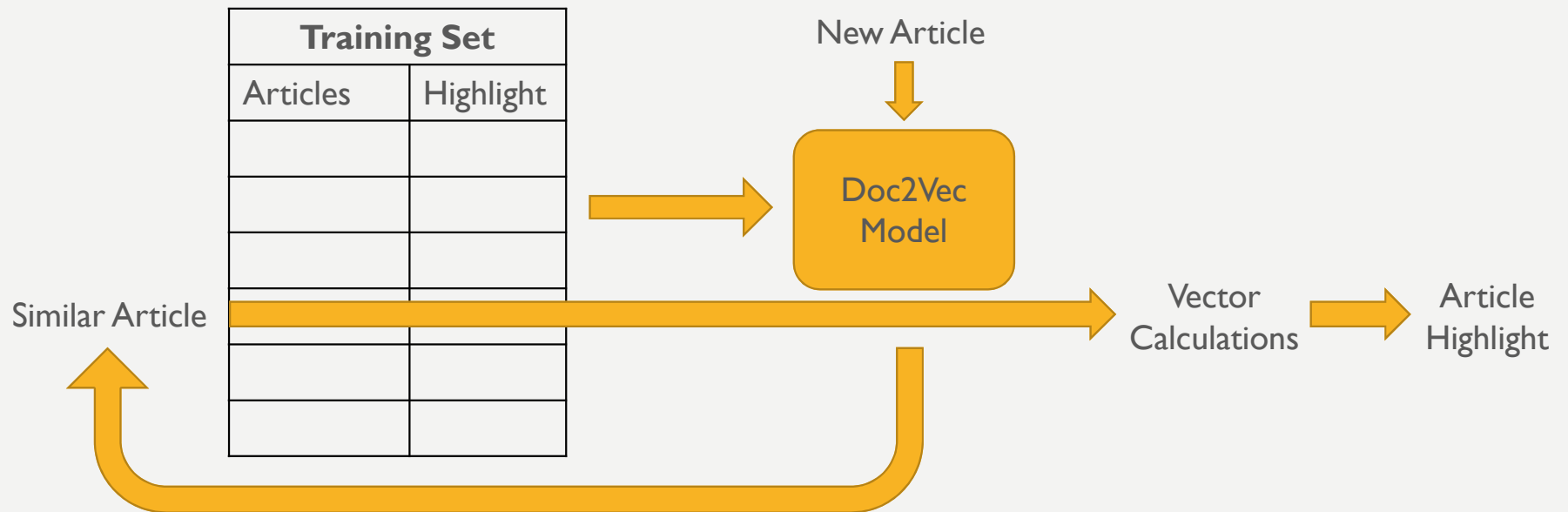
- Using this vector, find the most similar Medium article and its respective highlight

APPLYING DOC2VEC



- Apply vector calculations akin to Word2Vec
- $\text{Article} + \text{Highlight} - \text{New Article} = \text{New Highlight}$

APPLYING DOC2VEC



- With those calculations, the new article's highlight is calculated
- That calculated vector is then compared against each paragraph in the new article, outputting the most similar one

SAMPLE SUCCESSES

- **3 Things That Will Always Stand in the Way of Your Goals**

- To me, the only things anybody has a right to complain about are things like their health. Short of the death of a loved one, or a terminal illness, or some other horrible tragedy, everything feels controllable. If you're in control of it, you have the ability to fix it. Where is the value in complaining? Instead of complaining, my process is this: **Assess the problem, find the solution, and get on the offense. I'm an offensive player; complaining is playing defense.**

SAMPLE MISSES

- **3 Pages Every Morning: Why I Started a Daily Ritual and How I Stuck With It**

– I’ve come to love my morning pages. Which is unlike my love/hate relationship with writing publicly (my soul seems to require it, but damnit the process feels like wrestling with the Hulk). I love my morning pages because there is no heaviness in the outcome of the pages. It’s a form of unnecessary creation—just writing for the sake of writing. **There may be heaviness in trying to form the habit of writing itself (more on this below), but once you crack that, it’s all lightness. It’s a lightness that lives in caring about the process itself, not in the outcome of that process.**

FUTURE IMPROVEMENTS

- Work on Flask App for general usage
- Improve results by collecting a larger training set
- Include a wider variety of articles

THANK YOU

 CONTACT@LJAMESHU.COM

 [LJAMESHU](#)

 [LJAMESHU](#)