

# **Product Ads Recommendation System for Web Pages**

Sanjana Maiya

# Motivation

The screenshot shows a web browser displaying a New York Times article. The browser's address bar shows the URL: [www.nytimes.com/2015/06/05/technology/apple-to-unveil-tool-kit-for-watch-app-developers.html?ref=technology&module=Ribbon&version=origin&region=Header&action=click&contentCollection=technology](http://www.nytimes.com/2015/06/05/technology/apple-to-unveil-tool-kit-for-watch-app-developers.html?ref=technology&module=Ribbon&version=origin&region=Header&action=click&contentCollection=technology). The page header includes the word "TECHNOLOGY" and the article title "Apple to Unveil Tool Kit for Watch App Developers". Below the header is a navigation bar with social media icons and a "MORE" link. The main content area features a large image of a person working on a device, with the text "Apple to Unveil Tool Kit for Watch App Developers" overlaid. To the right of the main image are four smaller article thumbnails: "STATE OF THE ART Spotify Sees a Future Where Music Genres Don't Really Matter", "Apple's Foray Into Streaming Music Isn't a Guaranteed Success", "APP SMART Video Feature: Streaming Music Apps for a Variety of Tastes", and "ROBOTICA EPISODE 4 A Talking Teddy Bear Practicing in the Pediatric Hospital". The main text of the article begins with "anything really impressive," Mr. Ryu said about apps for the watch, which was released with around 3,500 programs. Now, "it sounds like they plan on untying our hands." The text continues: "Gaining more apps for Apple Watch is crucial to [Timothy D. Cook](#), Apple's chief executive. The device is his first new product since taking the reins of the company in 2011. Apps from outside developers helped increase sales of Apple's iPhones and iPads by vastly expanding their capabilities. Likewise, third-party apps could become important selling points for the watch, which now is primarily an expensive timepiece that requires an [iPhone](#) to operate. Apart from more powerful watch apps, Apple's developer conference is also set to showcase software advancements for OS X, the Mac operating system, and iOS, the mobile operating system powering iPhones and iPads. In addition, Apple plans to unveil a [new streaming music service](#) it developed with Beats, the music company it acquired for \$3 billion last year, according to people briefed on the situation, who declined to be identified because the details were confidential. Yet one much ballyhooed device will be absent from the conference: a new Apple TV, Apple's set-top box for televisions. The company planned as

anything really impressive," Mr. Ryu said about apps for the watch, which was released with around 3,500 programs. Now, "it sounds like they plan on untying our hands."

Gaining more apps for Apple Watch is crucial to [Timothy D. Cook](#), Apple's chief executive. The device is his first new product since taking the reins of the company in 2011. Apps from outside developers helped increase sales of Apple's iPhones and iPads by vastly expanding their capabilities. Likewise, third-party apps could become important selling points for the watch, which now is primarily an expensive timepiece that requires an [iPhone](#) to operate.

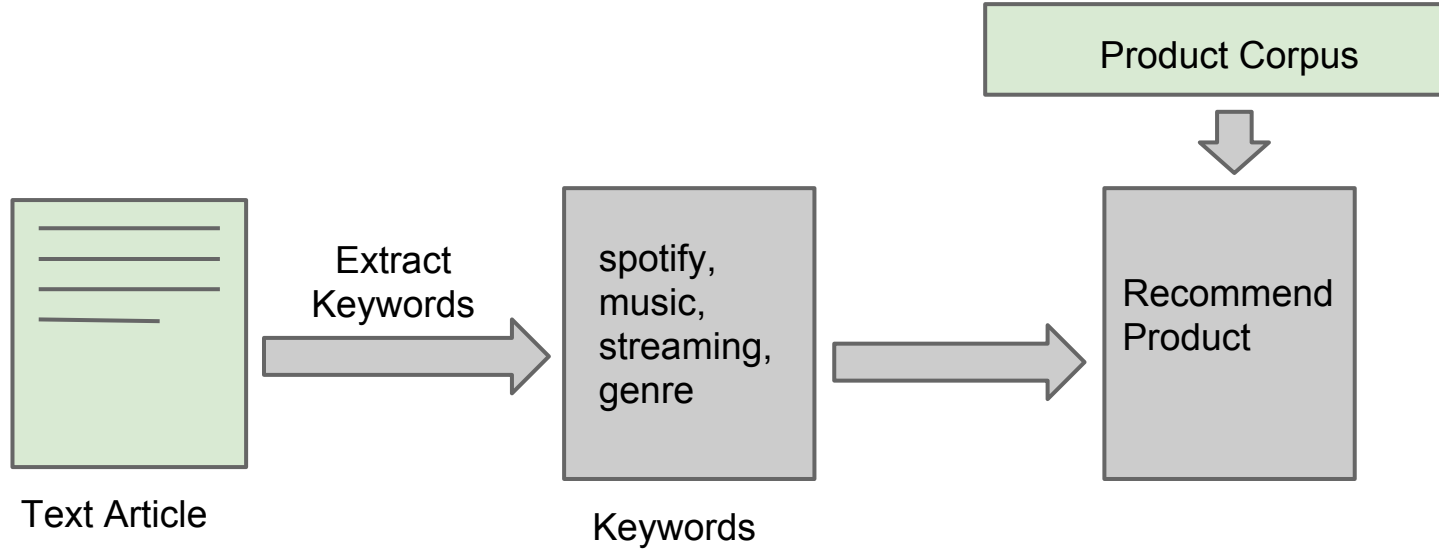
Apart from more powerful watch apps, Apple's developer conference is also set to showcase software advancements for OS X, the Mac operating system, and iOS, the mobile operating system powering iPhones and iPads. In addition, Apple plans to unveil a [new streaming music service](#) it developed with Beats, the music company it acquired for \$3 billion last year, according to people briefed on the situation, who declined to be identified because the details were confidential.

Yet one much ballyhooed device will be absent from the conference: a new Apple TV, Apple's set-top box for televisions. The company planned as

SAVOR ONE RARE, SMALL-LOT COFFEE EACH MONTH

Roasted in Seattle,

# High-level Design



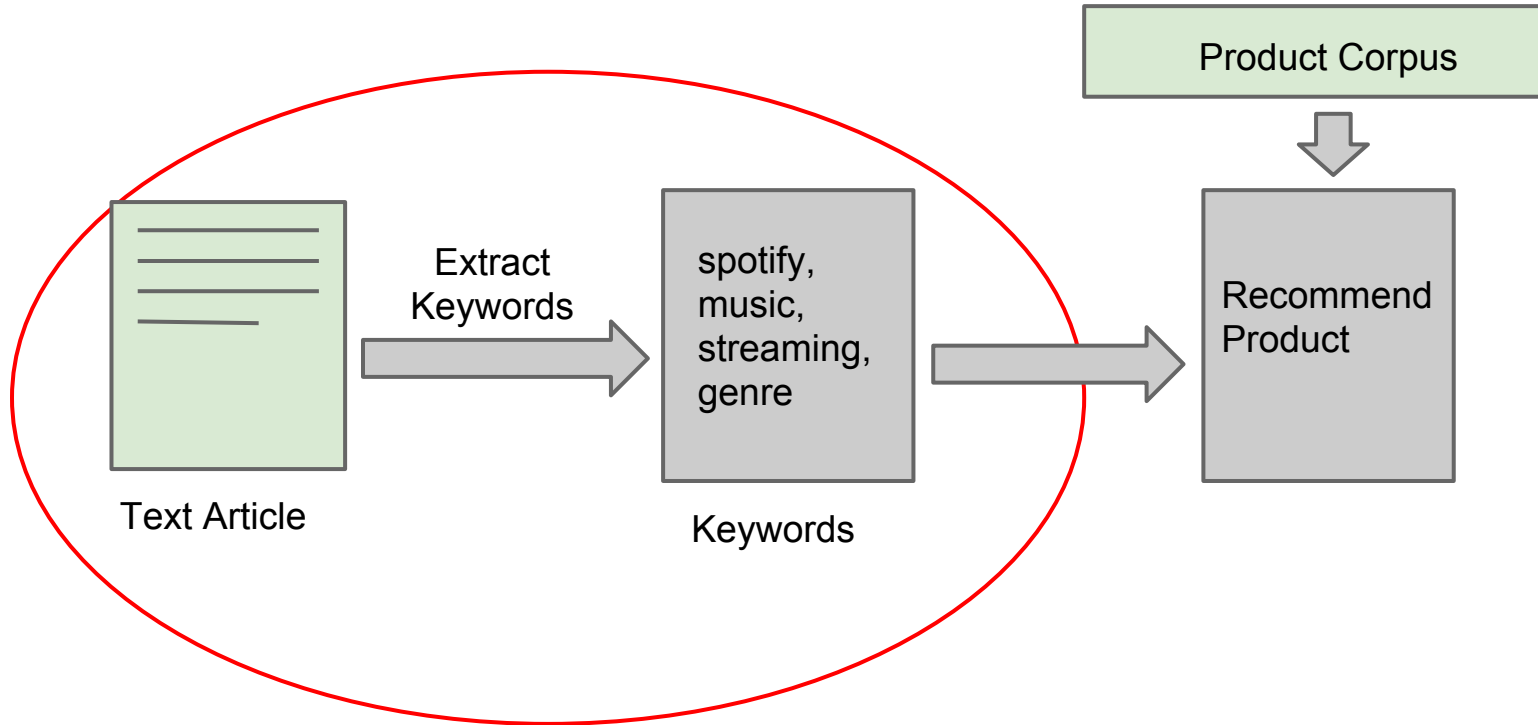
# Datasets

- Amazon products
  - 19 mil product metadata
- BBC news articles
  - 2200 news articles, unlabelled.

- Step 1: Keyword extraction.
- Step 2: Product recommendation.

- Step 1: Keyword extraction.
- Step 2: Product recommendation.

# Keyword Extraction



# Keyword Extraction Approaches

- Unsupervised Approach
  - Heuristic formula
- Supervised machine learning
  - Use training data to learn



# Keyword Extraction

- Built on top of RAKE, the Rapid Automatic Keyword Extraction Algorithm
- Score calculated based on
  - term frequency
  - phrase length
  - part-of-speech tagging
  - location of keywords in the document. This feature was dropped since it was adding noisy words.

# Keyword Extraction Example

## **Musical treatment for Capra film**

The classic film *It's A Wonderful Life* is to be turned into a musical by the producer of the controversial hit show *Jerry Springer - The Opera*.

Frank Capra's 1946 movie starring James Stewart, is being turned into a £7m musical by producer Jon Thoday. He is working with Steve Brown, who wrote the award-winning musical *Spend Spend Spend*. A spokeswoman said the plans were in the "very early stages", with no cast, opening date or theatre announced.

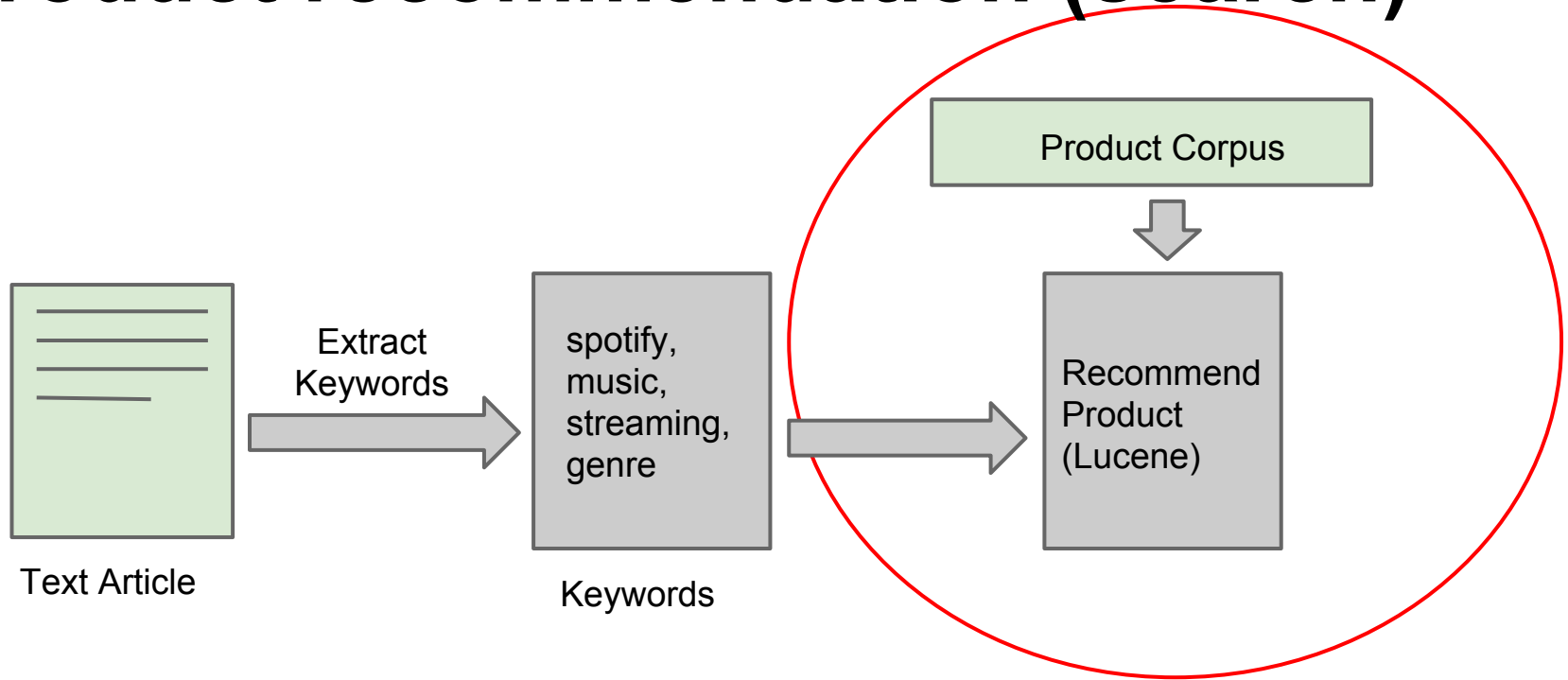
A series of workshops have been held in London, and on Wednesday a cast of singers unveiled the musical to a select group of potential investors. Mr Thoday said the idea of turning the film into a musical had been an ambition of his for almost 20 years. *It's a Wonderful Life* was based on a short story, *The Greatest Gift*, by Philip van Doren Stern. Mr Thoday managed to buy the rights to the story from Van Doren Stern's family in 1999, following Mr Brown's success with *Spend Spend Spend*. He later secured the film rights from Paramount, enabling them to use the title *It's A Wonderful Life*.

## **Keywords :**

1. **controversial hit show Jerry Springer**
2. **Philip van Doren Stern**

- Step 1: Keyword extraction.
- Step 2: Product recommendation.

# Product recommendation (search)



# Product recommendation (search)

- Around 19.4 mil Amazon products are indexed using Lucene
- Search for a product is done using keywords extracted
- Scoring is based on
  - tf, idf, article length
  - pricing information of product was not used since it reduced relevance of the retrieved ads

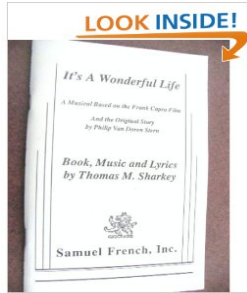
# Product recommendation (search)

- If score is below a certain threshold, no product ads displayed
- Word list maintained to prevent product ads being shown for sensitive content in news article

# Results

**keywords** : controversial hit show Jerry Springer Philip van Doren Stern

**Top products recommended:**



It's a wonderful life: a new musical based on the frank capra film and the original story by philip van doren stern

The greatest gift by philip van doren stern  
blue limited edition reproduction by graphic  
image



# Results

## Microsoft seeking spyware trojan

Microsoft is investigating a trojan program that attempts to switch off the firm's anti-spyware software.

The spyware tool was only released by Microsoft in the last few weeks and has been downloaded by six million people. Stephen Toulouse, a security manager at Microsoft, said the malicious program was called Bankash-A Trojan and was being sent as an e-mail attachment. Microsoft said it did not believe the program was widespread and recommended users to use an anti-virus program. The program attempts to disable or delete Microsoft's anti-spyware tool and suppress warning messages given to users.

It may also try to steal online banking passwords or other personal information by tracking users' keystrokes.

Microsoft said in a statement it is investigating what it called a criminal attack on its software. Earlier this week, Microsoft said it would buy anti-virus software maker Sybari Software to improve its security in its Windows and e-mail software. Microsoft has said it plans to offer its own paid-for anti-virus software but it has not yet set a date for its release. The anti-spyware program being targeted is currently only in beta form and aims to help users find and remove spyware - programs which monitor internet use, causes advert pop-ups and slow a PC's performance.



# Results

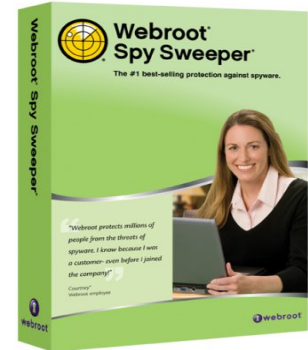
**Keywords :** virus software maker Sybari Software online banking passwords  
Microsoft spyware trojan

**Top products recommended :**



iyogi online virus and spyware removal

spy sweeper anti-spyware



# Results

## **Labour chooses Manchester**

The Labour Party will hold its 2006 autumn conference in Manchester and not Blackpool, it has been confirmed.

The much trailed decision was ratified by Labour's ruling National Executive Committee in a break with the traditional choice of a seaside venue. It will be the first time since 1917 that the party has chosen Manchester to host the annual event. Blackpool will get the much smaller February spring conference instead in what will be seen as a placatory move.

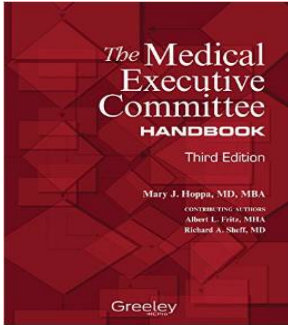
For years the main political parties have rotated between Blackpool, Bournemouth and Brighton. And the news the much larger annual conference is not to gather in Blackpool will be seen as a blow in the coastal resort. In 1998 the party said it would not return to Blackpool but did so in 2002. The following year Bournemouth hosted the event before the party signed a two year deal for Brighton to host the autumn conference.

Colin Asplin, Blackpool Hotel Association said: "We have tried very hard to make sure they come back to Blackpool. "Obviously we have failed in that. I just hope Manchester can handle the crowds. "It amazes me that the Labour Party, which is a working class party, doesn't want to come to the main working class resort in the country." The exact cost to Blackpool in terms of lost revenue for hotel accommodation is not yet known but it is thought that block bookings will be taken at the major Manchester hotels after the official announcement.

# Results

**keywords** : ruling National Executive Committee smaller February spring conference

**Top products recommended:**



medical executive committee handbook

national geographic february 1949



# Evaluation

## Mechanical Turk to Evaluate

- Keyword Extraction
- Ad relevance



# Keyword Evaluation

- Around 1000 keywords annotated using Mechanical Turk, however the quality of the keywords was poor
- For a test with 100 articles, the Precision of keywords extracted was 13.22%, Recall 18.37%, f1 measure 0.15.
- Decide to find how well our ad system did end-to-end versus retrieval of ads using the labelled keywords.

# End-to-end Evaluation

- First baseline was ad retrieval using title of articles as keywords
- Second baseline was ad retrieval using labelled keywords
- Both of the above were compared to the end-to-end system

# End-to-end Evaluation

## Metrics Used:

- **Coverage** : The total number of articles for which ads were recommended. This metric indicates the ability of the system to recommend ads. Since the system recommends up to 3 ads, the ad coverages for first, second and third ad are considered. For example, if the Ad coverage for the second ad is 65%, it means that we retrieved the second ad for only 65% of news articles.
- **Precision** : Amongst the ads that were displayed, the percentage of ads that were relevant. This evaluation is done for the first, second and third ad recommended. Overall precision is calculated over all three ads, where  $\text{overall precision} = (\text{number of relevant ads displayed} / \text{total number of ads displayed})$

Ad Coverage for Ad 1	100%
Ad Coverage for Ad 1	98%
Ad Coverage for Ad 1	98%
Precision for Ad 1	30%
Precision for Ad 2	19.38%
Precision for Ad 3	18.36%
Overall Precision	22.63%

Table 1:  
Ad Coverage and Precision for  
Baseline 1 (title as keywords)

Ad Coverage for Ad 1	79%
Ad Coverage for Ad 1	65%
Ad Coverage for Ad 1	56%
Precision for Ad 1	63.29%
Precision for Ad 2	49.23%
Precision for Ad 3	55.35%
Overall Precision	56.5%

Table 2:  
Ad Coverage and Precision for  
Baseline 2 (labelled keywords)



Ad Coverage for Ad 1	73%
Ad Coverage for Ad 1	58%
Ad Coverage for Ad 1	45%
Precision for Ad 1	78.08%
Precision for Ad 2	79.31%
Precision for Ad 3	77.77%
Overall Precision	78.4%

Table 3:  
Ad Coverage and Precision for  
our ad system

- Our ad system does much better than both baselines in terms of precision of ads retrieved
- Clearly, the quality of the keywords annotated was not up to the mark
- Coverage has dropped slightly in our system, which is not necessarily a bad thing, since it is better not to display ads than to show ads which are completely irrelevant

# Conclusion

- Evaluation of relevance of Ads not straightforward
- May not be always possible to display relevant ad
- A lot of scope for extension
  1. User purchasing history for Amazon
  2. Personalized Ads based on browsing history/logs
  3. User similarity measures

# Future Direction

- Using ratings to recommend Products
- Improving the keyword extraction process, more accurate labelling of keywords so that learning can be done for keyword extraction
- Tuning the search based on user feedback

# References

1. BBC dataset: D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006.
2. Amazon dataset : Image-based recommendations on styles and substitutes J. McAuley, C. Targett, J. Shi, A. van den Hengel SIGIR, 2015
3. RAKE algorithm : Rapid Automatic Keyword Extraction (RAKE) algorithm as described in: Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010)
4. POStagger : D. Roth and D. Zelenko, Part of Speech Tagging Using a Network of Linear Separators. Coling-Acl, The 17th International Conference on Computational Linguistics (1998)
5. Finding Advertising Keywords on Web Pages - Wen-tau Yih, Joshua Goodman, Vitor R. Carvalho, WWW 2006