# Basketball Article Recommender for Bleacher Report

**Stephen Albro**
Master of Business Analytics
Massachusetts Institute of Technology
Cambridge, MA 02139
salbro@mit.edu

**Cyrille Combettes**
Master of Business Analytics
Massachusetts Institute of Technology
Cambridge, MA 02139
cyrille@mit.edu

## 1 Abstract and Motivation

As avid basketball fans, we needed to keep up-to-date on the National Basketball Association during our busy MBAn year. Bleacher Report, a primary digital destination for sports readers, kept us up to speed all year with its well-written articles, and so we thought about returning the favor in a project.

In an age of cut-throat digital competition, it is vital for websites to retain visitors. Bleacher Report's hope for a visitor is two-fold. First, that the visitor jumps from article to article on the website, rather than returning to the search engine results page. Second, that the visitor falls in love with the content and coverage of Bleacher Report's articles, and develops a loyalty to the website.

We feel that having a micro-targeted article recommendation system would provide Bleacher Report with an analytics-based edge over its competitors. In this project, we developed a prototype for an article recommendation system using state of the art NLP techniques. Using our best model, we also produced visualizations of the Bleacher Report topic landscape and discovered interesting surprises.

## 2 Data collection and preprocessing

We queried Bleacher Report NBA articles from October 17, 2017, the start of the 2017-2018 NBA season. Our entire query consisted of joining words like *basketball* and *NBA* with all 30 teams and the top 25 active players: we received 3,314 URLs in return. We then wrote a script to request the webpage for each URL and scrape its content, picking the paragraphs out of the HTML.

There were a number of preprocessing steps we had to do before our articles were ready for analysis. Because our primary goal was text-based clustering, a bag-of-words cleaning approach was preferred over one that preserved punctuation and grammar.

## 3 Key results

A reasonable document-recommendation strategy embed documents as vectors and then use K Nearest Neighbors to group documents together. The first step in our project, then, was to choose an article embedding strategy. We experimented with three primary methods: word2vec, Non-negative Matrix Factorization (NMF), and Latent Dirichlet Allocation (LDA). The later two methods discover underlying *topics* in the articles and then express each document as a distribution over those topics. We evaluated each approach using a hand-crafted test set.

A preliminary analysis revealed LDA to be the most promising strategy, and so we focused the rest of our efforts on developing a quality LDA topic model, which we then use to produce visualizations of the topic landscape of Bleacher Report's NBA articles. To our surprise, the final clusters revealed that a couple of other major sports leagues snuck into our primarily NBA-driven corpus ...

15.071C Advanced Analytics Edge: project abstract. May 10, 2018. Massachusetts Institute of Technology, MA.