

# Finding Duplicate News Articles

A New Way of Shingling  
Bucketing by Length

Mining of Massive Datasets  
Leskovec, Rajaraman, and Ullman  
Stanford University



# Application: Same News Article

- The Political-Science Dept. at Stanford asked a team from CS to help them with the problem of identifying duplicate, on-line news articles.
- **Problem:** the same article, say from the Associated Press, appears on the Web site of many newspapers, but looks quite different.

# News Articles – (2)

- Each newspaper surrounds the text of the article with:
  - It's own logo and text.
  - Ads.
  - Perhaps links to other articles.
- A newspaper may also “crop” the article (delete parts).

# News Articles – (3)

- The team came up with its own solution, that included shingling, but not minhashing or LSH.
- A special way of shingling that appears quite good for **this** application.
- **LSH substitute**: candidates are articles of similar length.

# Enter LSH

- I told them the story of minhashing + LSH.
- They implemented it and found it faster for similarities below 80%.
  - **Aside:** That's no surprise. When similarity is high, there are better methods, as we shall see.

# Enter LSH – (2)

- Their first attempt at minhashing was very inefficient.
- They were unaware of the importance of doing the minhashing row-by-row.
- Since their data was column-by-column, they needed to sort once before minhashing.

But the problem was that I forgot to remind them to do the minhashing row by row, where you compute the hash value for each row number once and for all rather than once for each column. Remember that the rows correspond to the shingles and the columns to the web pages.

# Specialized Shingling Technique

- The team observed that news articles have a lot of *stop words*, while ads do not.
  - “Buy Sudzo” vs. “I recommend *that you* buy Sudzo *for your* laundry.”
- They defined a *shingle* to be a stop word and the next two following words.

# Why it Works

- By requiring each shingle to have a stop word, they biased the mapping from documents to shingles so it picked more shingles from the article than from the ads.
- Pages with the same article, but different ads, have higher Jaccard similarity than those with the same ads, different articles.