

Finding Similar Sets

Applications

Shingling

Minhashing

Locality-Sensitive Hashing

Mining of Massive Datasets

Leskovec, Rajaraman, and Ullman

Stanford University



Applications of Set-Similarity

Many data-mining problems can be expressed as finding “similar” sets:

1. Pages with similar words, e.g., for classification by topic.
2. Netflix users with similar tastes in movies, for recommendation systems.
3. **Dual**: movies with similar sets of fans.
4. Entity resolution.

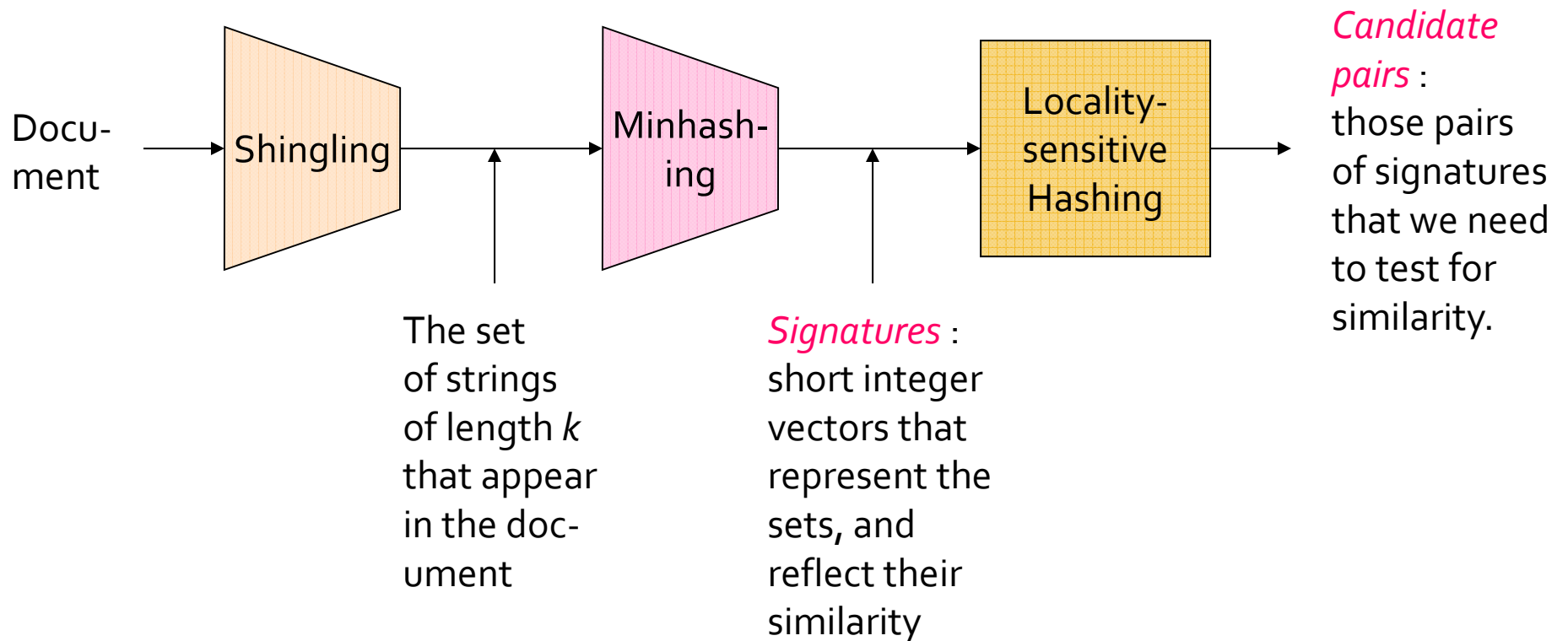
Similar Documents

- Given a body of documents, e.g., the Web, find pairs of documents with a lot of text in common, such as:
 - Mirror sites, or approximate mirrors.
 - **Application:** Don't want to show both in a search.
 - Plagiarism, including large quotations.
 - Similar news articles at many news sites.
 - **Application:** Cluster articles by “same story.”

Three Essential Techniques for Similar Documents

1. *Shingling* : convert documents, emails, etc., to sets.
2. *Minhashing* : convert large sets to short signatures, while preserving similarity.
3. *Locality-sensitive hashing* : focus on pairs of signatures likely to be similar.

The Big Picture



Shingles

- A k -shingle (or k -gram) for a document is a sequence of k characters that appears in the document.
- **Example:** $k=2$; doc = abcab. Set of 2-shingles = {ab, bc, ca}.
- Represent a doc by its set of k -shingles.

Shingles and Similarity

- Documents that are intuitively similar will have many shingles in common.
- Changing a word only affects k -shingles within distance k from the word.
- Reordering paragraphs only affects the $2k$ shingles that cross paragraph boundaries.
- **Example:** $k=3$, “The dog which chased the cat” versus “The dog that chased the cat”.
 - Only 3-shingles replaced are g_w , $_wh$, whi , hic , ich , $ch_$, and h_c .

Shingles: Compression Option

- To compress long shingles, we can hash them to (say) 4 bytes.
 - Called *tokens*.
- Represent a doc by its tokens, that is, the set of hash values of its k -shingles.
- Two documents could (rarely) appear to have shingles in common, when in fact only the hash-values were shared.