Information Retrieval
Spelling Correction using K-gram Overlapping
Lecture-14&15

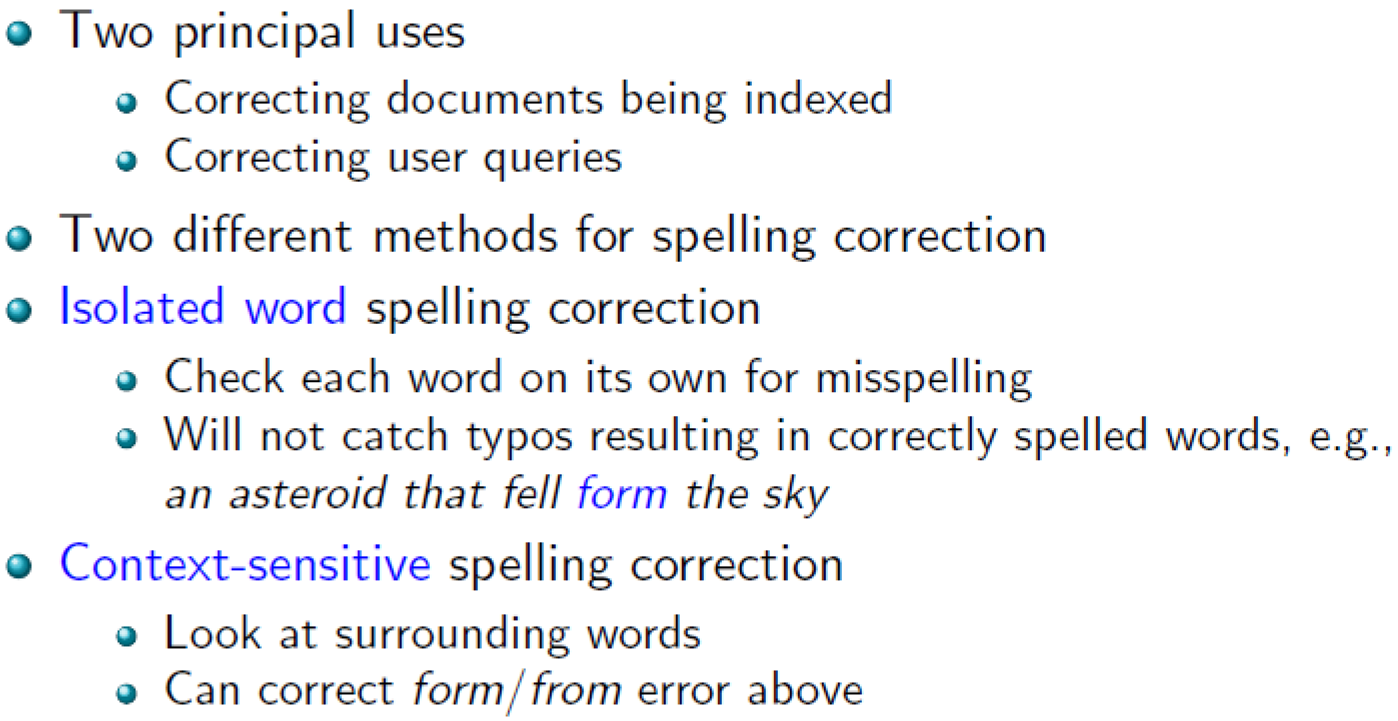
Prepared By

Dr. Rasmita Dash & Dr. Rasmita Rautray
Associate Professor
Dept. of CSE

Content

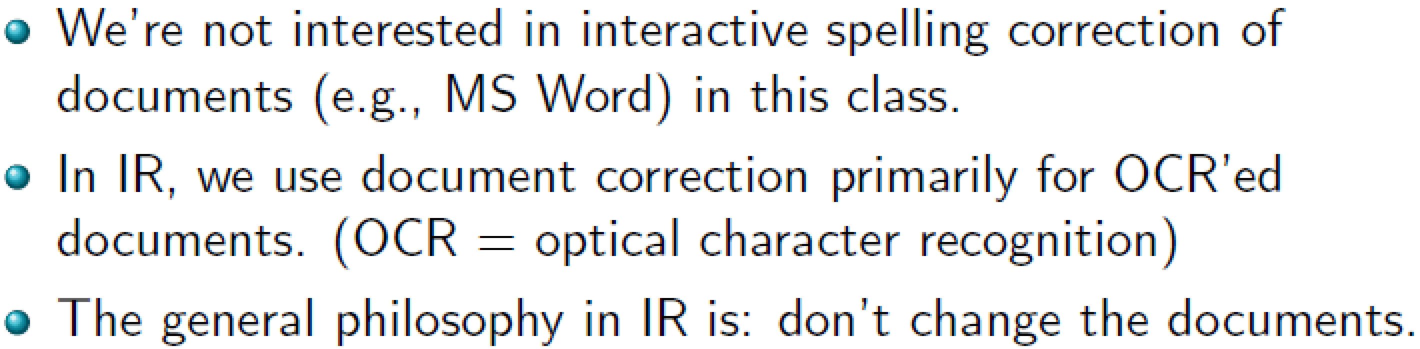
- Spelling correction
- Correcting Documents
- K gram for spelling correction
- Jaccard Coefficient

Spelling correction



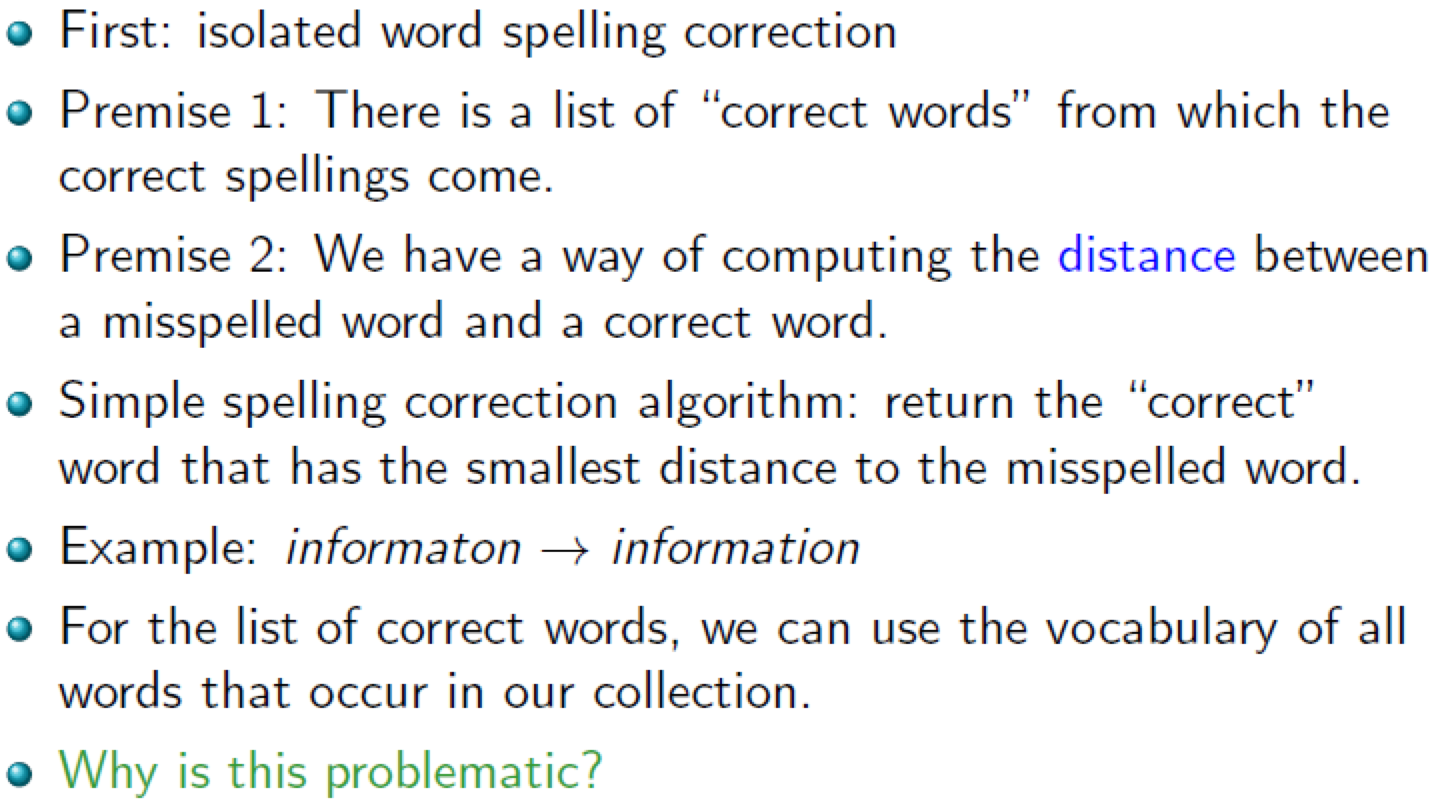
- Two principal uses
  - Correcting documents being indexed
  - Correcting user queries
- Two different methods for spelling correction
- Isolated word spelling correction
  - Check each word on its own for misspelling
  - Will not catch typos resulting in correctly spelled words, e.g., *an asteroid that fell form the sky*
- Context-sensitive spelling correction
  - Look at surrounding words
  - Can correct *form*/*from* error above

Correcting Documents



- We're not interested in interactive spelling correction of documents (e.g., MS Word) in this class.
- In IR, we use document correction primarily for OCR'ed documents. (OCR = optical character recognition)
- The general philosophy in IR is: don't change the documents.

Contd..



- First: isolated word spelling correction
- Premise 1: There is a list of “correct words” from which the correct spellings come.
- Premise 2: We have a way of computing the distance between a misspelled word and a correct word.
- Simple spelling correction algorithm: return the “correct” word that has the smallest distance to the misspelled word.
- Example: *informaton* → *information*
- For the list of correct words, we can use the vocabulary of all words that occur in our collection.
- Why is this problematic?

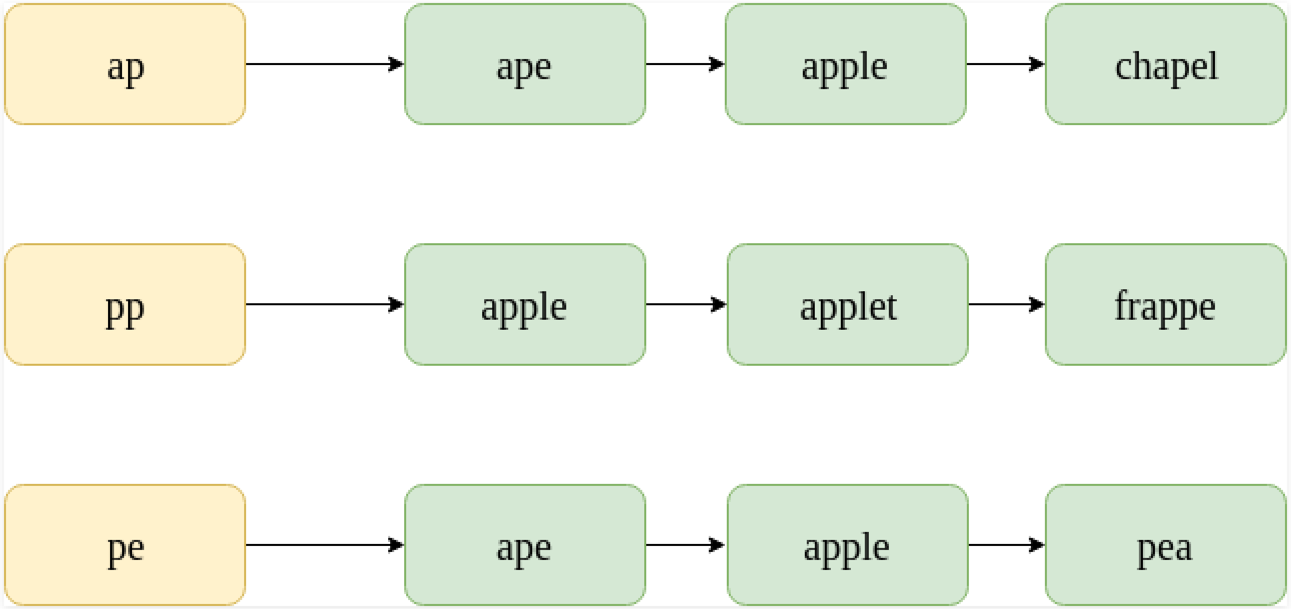
K gram for spelling correction

- The steps involved for spelling correction are:
- Find the k-grams of the misspelled word.
- For each k-gram, linearly scan through the postings list in the k-gram index.
- Find k-gram overlaps after having linearly scanned the lists (no extra time complexity because we are finding the Jaccard coefficient).
- Return the terms with the maximum k-gram overlaps.

Example

Consider the misspelt word: “appe”.

K gram index k=2.



ap → ape → apple → chapel

pp → apple → applet → frappe

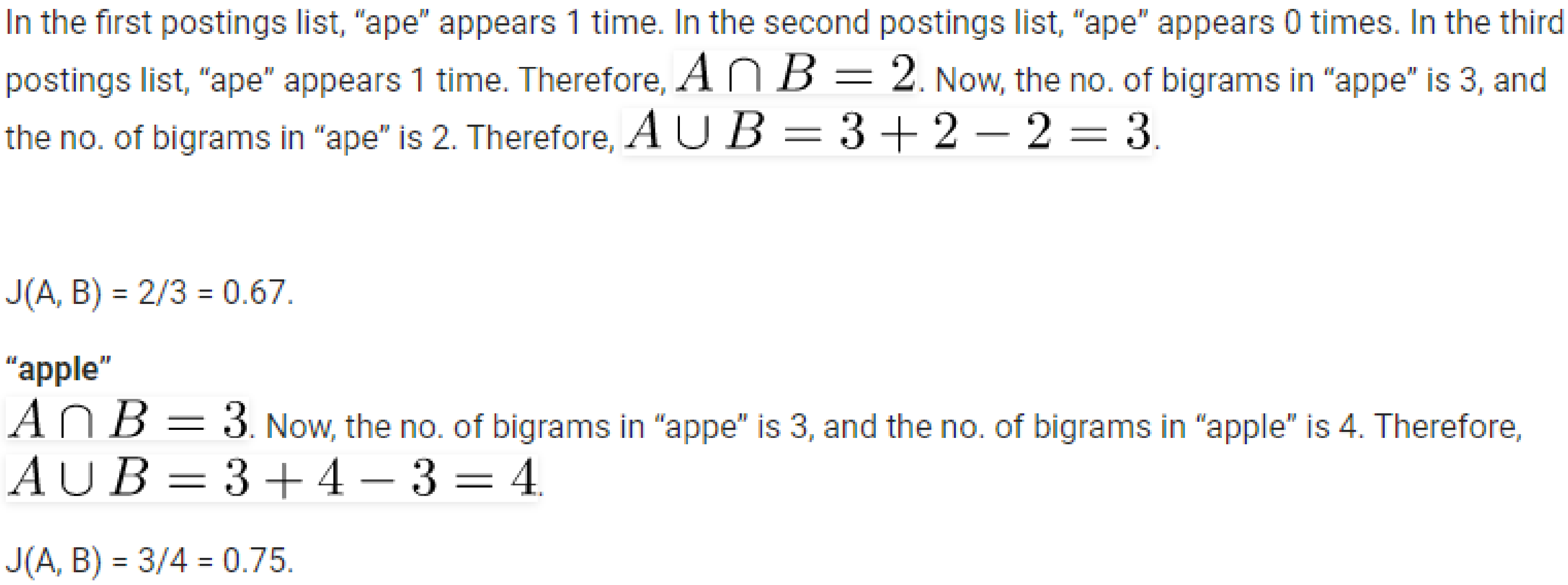
pe → ape → apple → pea

Jaccard Coefficient

To find the k-gram overlap between two postings list, we use the Jaccard coefficient. Here, A and B are two sets (postings lists), A for the misspelt word and B for the corrected word.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Candidate terms for spelling correction, namely “ape” and “apple”.



In the first postings list, “ape” appears 1 time. In the second postings list, “ape” appears 0 times. In the third postings list, “ape” appears 1 time. Therefore, $A \cap B = 2$. Now, the no. of bigrams in “appe” is 3, and the no. of bigrams in “ape” is 2. Therefore, $A \cup B = 3 + 2 - 2 = 3$.

J(A, B) = 2/3 = 0.67.

“apple”

$A \cap B = 3$. Now, the no. of bigrams in “appe” is 3, and the no. of bigrams in “apple” is 4. Therefore, $A \cup B = 3 + 4 - 3 = 4$.

J(A, B) = 3/4 = 0.75.

Thank You