IR vs NLP
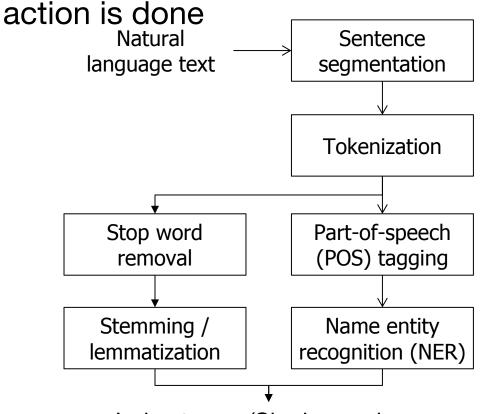
- Both IR and NLP deal with natural language texts
- IR: indexing, stop word, stemming, root word, …
- NLP: tokenization, lemmatization, morphological analysis, …
- IR is more efficiency minded, while NLP is semantics minded
- Their boundary is blurred: IR is increasingly semantics minded

The Document Pipeline

- There are different designs about which action should be done first, and which should be second, and where an action is done

Natural language text → **Sentence segmentation** ← Break a document into sentences

↓

Tokenization ← Convert a sentence into words

↓

Stop word removal | **Part-of-speech (POS) tagging** ← Identify syntactic classes for each token/word

↓ | ↓

Stemming / lemmatization | **Name entity recognition (NER)** ← Identify compound nouns, places, person names, etc.

↓

Index terms (Single words or phrases)