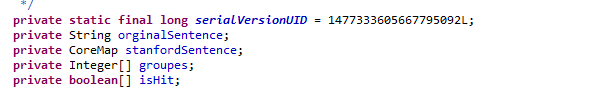
1. LGT Sentence (=unit of analysis) contains Coremap as a property (amongst others)

All feature extraction jobs run on sentences

Contains a number of generic operations on sentences (navigating within sentence such as token count between two reg expression)

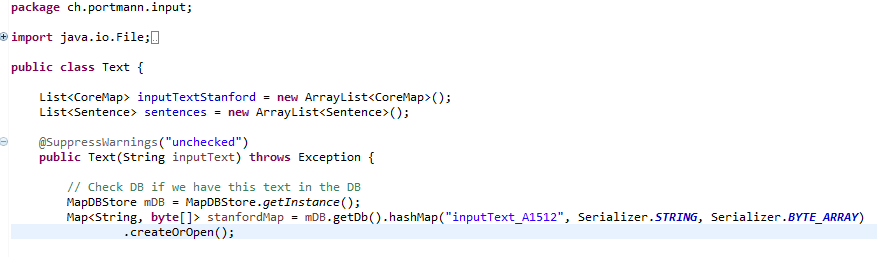
CoreMap, groups and isHit have same length = number of tokens in sentence



1. Serialize Coremap Annotations in constructor of Text

Hashmap (key = original sentence, value = Stanford annotation)

We serialize keys (Serializer.String) and values (Serializer.ByteArray) since hashmap works on serialized inputs only



1. CorpusRun = main

Instantiates ProcessFeature class for every file in corpus.

Files are “boilerpiped” raw \*.htmls

1. ProcessCompare and ProcessFeature

builds *compare* or *feature extraction* pipeline in method ProcessFeature.

Uses FeaturePipe: generic pipeline class implementing the LGTFeature interface (defined as containing a run(Sentence) method.