Peter El Jiz
6250661

	Distinct Terms			Non Positional Postings		
	#	Δ	T	#	Δ	T
Unfiltered	112420			1912577		
No Punctuation, No dash	54008	-51.96	-51.96	1871225	-2.16	-2.16
No Numbers	49949	-7.51	-55.57	1676335	-10.41	-12.35
Case Folding	43072	-13.76	-61.69	1620641	-3.32	-15.26
Stop Words	42779	-0.68	-61.94	1140096	-29.65	-40.38
Stemming	32404	-24.25	-71.17	1118817	-1.86	-41.5

SPIMI (1): Writing to files

- SPIMI is allocated x% of memory on parsing files and adding doc – term values to a list.
- It is allocated the remaining memory for storing all pairs sharing the same terms in postings list and storing the end result in a hash structure:
  Hash<String, List<Integer>>
  It then serializes (encodes as an object) the hash to disk.

SPIMI (2): Read and Merge

- The n blocks are collected. (0 and 1) are merged and serialized to disk as block0-1. Block 0-1 is then merged with Block 2 and etc
  merge ((k and k+1) and k + 2) and k + ... + n.

Query Processing and Retrieval

- The system first processes the query. (The query string is filtered the same way documents are, punctuation, dashes, numbers, case fold, stop words, and stem the words)
  AND: query method retrieves each word's posting list and intersects them.
  OR: add all docs from each word's posting, remove duplicates
- Retrieve documents and display them