



Team 4

Attack Westminster

Member: Colm Gallagher, Jamie Dyer,
Jeanine Liebold, Limin Yang

<https://github.com/whyisyoung/AttackWestminster>



Solr Index

- Created solr index (big_team4) from the Big Attack Westminster warc file
- 'test_team4' core still has the Hurricane Sandy data set
- Created json for Small Westminster Attack warc

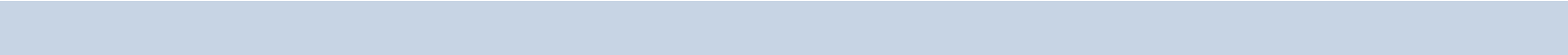


Statistics

Last Modified:	5 minutes ago
Num Docs:	11298
Max Doc:	11298
Heap Memory	-1
Usage:	
Deleted Docs:	0
Version:	14
Segment Count:	1
Current:	✓



PySpark Migration

- Migrated some scripts to utilize PySpark's distributed clustering capabilities
 - We were unable to get scripts to execute on the Hadoop cluster without using the `pyspark2` terminal. If other teams have run into this and found solutions, please let us know
- 

Data Cleaning

- Manually read dataset and find what maybe noisy
- Apply jusText recommended by Team 14
- Seems efficient, but needs more investigation
- Any recommendation for performance measurement?

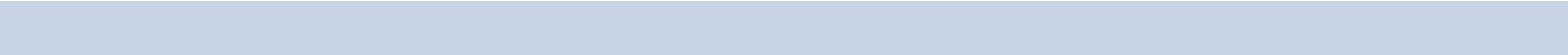
[illegible]

jusText

(accessed August 9, 2013) (accessed August 9, 2013)



Strategy for customized stopword

1. Google the event and create an approximate summary of it
 2. Expanding the list of customized stopword based on our summary
- 



Next Steps

- Replicate previous work on the new small dataset [Attack Westminster]
 - Use lemmatization to merge some of the results of Unit 2
 - Conclude the results of Unit 1 and 2 for the small dataset
 - PoS tagging (Unit 3)
- 