Basic Data

Number of Files	46
Number of Words	6793
Unique Words	1280
Avg Words/File	147.67
Avg Unique Words/File	27.83
Total Number of Items created in SDB	105

Current SDB limits are:

- 250,000,000 items / domain (they have stated they intent to increase this 4X in the near future)
- 256 unique attribute/value pairs per item
- 1024 bytes per attribute name/ value

Given even the current limits, I don't think scale would be a problem, especially if a separate domain was created per corpus.

Indexing Approach

Item Name is Doc ID + page number	Attribute Name is the word	Attribute Values are word locations within Doc ID
e73a388c-b8a8-40c5-a388-176bf1e656ac	pages	1
		2
e73a388c-b8a8-40c5-a388-176bf1e656ac:1	appointments	50
		65
		118
		131
		215
		234
	...	...
e73a388c-b8a8-40c5-a388-176bf1e656ac:2	appointments	296
	...	...

- Because of limit of 256 attribute/name pairs per item in SDB, indexes need to be split into pages.
- An item is stored using the Doc ID as the item name. This item is used to keep track of all pages associated with that Doc ID.
- Item name is constructed by concatenating document ID with page number using a semi-colon as the separator
- Attribute names are the words that the document contains
- Attribute values are locations of the word within the document. These are stored as multi-valued attributes pairs
- Current implementation does not attempt to keep all indexes for a particular word on a single page. This could potentially be accomplished with additional up-front processing of the text but the value is questionable.
- Current implementation excludes common articles from the index
- Current implementation does not perform stemming so "appointment" and "appointments" are indexed separately. This could be easily remedied using nltk.

Possible Queries available in this approach:

All documents containing the word "appointments"

```
>>> c.query('sdbindex7', "['appointments'starts-with'']")
[u'ca2ab648-c3b8-4903-b058-c455e6e5fb59:1', u'e73a388c-b8a8-40c5-a388-176bf1e656ac:1', u'0b3abfc5-3204-479d-8fd8-59306ea46f62:1',
u'0b3abfc5-3204-479d-8fd8-59306ea46f62:6', u'e73a388c-b8a8-40c5-a388-176bf1e656ac:2', u'0b3abfc5-3204-479d-8fd8-59306ea46f62:2',
u'98d5af42-3c50-4f8a-8e97-9bf12da9b4ac:1', u'1e85b9c0-ce61-49b9-9f0b-47ab19c44050:1', u'e73a388c-b8a8-40c5-a388-176bf1e656ac:4',
u'0b3abfc5-3204-479d-8fd8-59306ea46f62:4']
```

All documents containing the word "appointments" and the word "minnesota"

```
c.query('sdbindex7', "['appointments'starts-with'']AND['minnesota'starts-with'']")
[u'ca2ab648-c3b8-4903-b058-c455e6e5fb59:1']
```

Storage Cost Estimates

Total Number of Items:	105		
Total Number of Name/Value Pairs:	11274		
Total Storage (in bytes):	597126		
Estimated Storage Cost/Month:	$0.0027		
Projections (assumes sample data is representative)			
Monthly Cost / Document	$0.000058		
Monthly Cost for 1 million Documents:	$58.41		
Monthly Cost / Word:	$0.00000040		
Monthly Cost for 100 million Words:	$395.56		

- Storage Costs are $.0045 / MB / Month
- Storage calculated by counting byte size of each item and each unique name/value attribute pair and adding overhead of 45 bytes for each item and each unique name/value attribute pair