TextCritical.net - Bug #1258

Work indexer fails

03/11/2016 06:58 AM - Luke Murphey

Status:	Closed	**Start date:**	03/11/2016
Priority:	Normal	**Due date:**	
Assignee:	Luke Murphey	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	1.3.1		

Description

```
^[[C^[[CTraceback (most recent call last):
  File "manage.py", line 10, in <module>
    execute_from_command_line(sys.argv)
  File "/Library/Python/2.7/site-packages/django/core/management/__init__.py", line 354, in execut
e_from_command_line
    utility.execute()
  File "/Library/Python/2.7/site-packages/django/core/management/__init__.py", line 346, in execut
e
    self.fetch_command(subcommand).run_from_argv(self.argv)
  File "/Library/Python/2.7/site-packages/django/core/management/base.py", line 394, in run_from_a
rgv
    self.execute(*args, **cmd_options)
  File "/Library/Python/2.7/site-packages/django/core/management/base.py", line 445, in execute
    output = self.handle(*args, **options)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/management/commands/mak
e_search_indexes.py", line 36, in handle
    WorkIndexer.index_all_works()
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/contentsearch.py", line
 124, in index_all_works
    cls.index_work(work)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/contentsearch.py", line
 164, in index_work
    cls.index_division(division)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/contentsearch.py", line
 187, in index_division
    writer.commit()
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/writing.py", line 922,
in commit
    finalsegments = self._merge_segments(mergetype, optimize, merge)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/writing.py", line 827,
in _merge_segments
    return mergetype(self, self.segments)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/writing.py", line 101,
in MERGE_SMALL
    writer.add_reader(reader)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/writing.py", line 709,
in add_reader
    docmap = self.write_per_doc(fieldnames, reader)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/writing.py", line 690,
in write_per_doc
    v = reader.vector(docnum, fieldname, fieldobj.vector)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/reading.py", line 832,
in vector
    return self._perdoc.vector(docnum, fieldname, vformat)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/codec/whoosh3.py", line
 482, in vector
    byteids=True)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/codec/whoosh3.py", line
 879, in __init__
    self._read_header()
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/codec/whoosh3.py", line
 890, in _read_header
    raise Exception("Block tag error %r" % magic)
Exception: Block tag error 'VPST'
```

The last log entry is:

```
reader.contentsearch: Successfully indexed verse, verse=4, division="36", work="Cynegeticus"
```

History

#1 - 03/11/2016 10:25 PM - Luke Murphey

Seems to be crashing after the following:

```
Successfully indexed verse, verse=4, division="36", work="Cynegeticus"
```

Tried to regenerate it with the following:

```
python manage.py make_search_indexes -c -w "Cynegeticus"
```

This worked however.

#2 - 03/11/2016 10:38 PM - Luke Murphey

Moving the index commit to the work level makes it faster.

#3 - 03/12/2016 07:09 AM - Luke Murphey

Now it is dying at:

```
Successfully indexed division, division="chapter 89", work="The Deipnosophists, Book 10"
```

#4 - 03/12/2016 07:10 AM - Luke Murphey

Consistently fails on:

```
python manage.py make_search_indexes -w the-deipnosophists-book-10
```

#5 - 03/12/2016 07:41 AM - Luke Murphey

Try indexing the work and committing by verse to identify the verse that is the problem.

#6 - 03/12/2016 05:53 PM - Luke Murphey

Observations:

- The indexing fails at the commit stage
- Once the error is observed, no more commits will work
- It is failing on the-deipnosophists-book-10
- Indexing the-deipnosophists-book-10 works after clearing the indexes; then indexing the-deipnosophists-book-11 works too as well as Cynegeticus
- Whoosh 2.4.1 seemed to index fine

#7 - 03/12/2016 05:54 PM - Luke Murphey

I could try indexing on Windows to see if it reproduces on other platforms.

#8 - 03/12/2016 06:30 PM - Luke Murphey

Trying different versions of Whoosh:

1. 2.4.1: works
2. 2.7.2: doesn't
3. 2.5.0: gets different error: IndexError: list index out of range after "reader.contentsearch: Successfully indexed verse, verse=3, division="35", work="Cynegeticus""

2.5.0 error:

```
Traceback (most recent call last):
  File "manage.py", line 10, in <module>
    execute_from_command_line(sys.argv)
  File "/Library/Python/2.7/site-packages/django/core/management/__init__.py", line 354, in execute_comma
nd_line
    utility.execute()
  File "/Library/Python/2.7/site-packages/django/core/management/__init__.py", line 346, in execute
    self.fetch_command(subcommand).run_from_argv(self.argv)
  File "/Library/Python/2.7/site-packages/django/core/management/base.py", line 394, in run_from_argv
    self.execute(*args, **cmd_options)
  File "/Library/Python/2.7/site-packages/django/core/management/base.py", line 445, in execute
    output = self.handle(*args, **options)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/management/commands/make_search_ind
exes.py", line 42, in handle
    WorkIndexer.index_all_works()
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/contentsearch.py", line 165, in ind
ex_all_works
    cls.index_work(work, commit=True)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/contentsearch.py", line 216, in ind
ex_work
    cls.index_division(division, commit=False, writer=writer)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/contentsearch.py", line 239, in ind
ex_division
    cls.index_verse(verse, division=division, writer=writer, commit=False)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/contentsearch.py", line 324, in ind
ex_verse
    author        = author_str
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/writing.py", line 752, in add_docum
ent
    perdocwriter.add_vector_items(fieldname, field, vitems)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/codec/whoosh3.py", line 232, in add
_vector_items
    vinfo = vpostwriter.finish_postings()
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/codec/whoosh3.py", line 648, in fin
ish_postings
    terminfo.add_block(self)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/codec/whoosh3.py", line 1083, in ad
d_block
    self._minid = block.min_id()
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/codec/whoosh3.py", line 776, in min_id
    return self._ids[0]
```

#9 - 03/12/2016 06:30 PM - Luke Murphey

Going to try disabling term vectors

#10 - 03/12/2016 06:38 PM - Luke Murphey

2.4.1 indexing fails:

```
Traceback (most recent call last):
  File "manage.py", line 10, in <module>
    execute_from_command_line(sys.argv)
  File "/Library/Python/2.7/site-packages/django/core/management/__init__.py", line 354, in execute_from_command_line
    utility.execute()
  File "/Library/Python/2.7/site-packages/django/core/management/__init__.py", line 346, in execute
    self.fetch_command(subcommand).run_from_argv(self.argv)
  File "/Library/Python/2.7/site-packages/django/core/management/base.py", line 394, in run_from_argv
    self.execute(*args, **cmd_options)
  File "/Library/Python/2.7/site-packages/django/core/management/base.py", line 445, in execute
    output = self.handle(*args, **options)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/management/commands/make_search_indexes.py", line 42, in handle
    WorkIndexer.index_all_works()
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/contentsearch.py", line 165, in index_all_works
    cls.index_work(work, commit=True)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/reader/contentsearch.py", line 219, in index_work
    writer.commit()
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/filedb/filewriting.py", line 502, in commit
    finalsegments = self._merge_segments(mergetype, optimize, merge)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/filedb/filewriting.py", line 432, in _merge_segments
    return mergetype(self, self.segments)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/filedb/filewriting.py", line 68, in MERGE_SMALL
    writer.add_reader(reader)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/filedb/filewriting.py", line 328, in add_reader
    self._merge_per_doc(reader, docmap)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/filedb/filewriting.py", line 309, in _merge_per_doc
    perdocwriter.add_vector_matcher(fieldname, field, v)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/codec/whoosh2.py", line 219, in add_vector_matcher
    self.add_vector_items(fieldname, fieldobj, readitems())
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/codec/whoosh2.py", line 209, in add_vector_items
    self.vindex.add((self.docnum, fieldname), startoffset)
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/filedb/filetables.py", line 490, in add
    self._add(self.keycoder(key), self.valuecoder(data))
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/filedb/filetables.py", line 122, in add
    self.add_all(((key, value),))
  File "/Users/lmurphey/Documents/SP/Workspace/TextCritical.com/src/whoosh/filedb/filetables.py", line 346, in add_all
    raise ValueError("Keys must increase: %r .. %r" % (lk, key))
ValueError: Keys must increase: '\x00\x00\x00^\x00\x01' .. '\x00\x00\x00^\x00\x00'
```

#11 - 03/12/2016 07:25 PM - Luke Murphey

Fails on Windows too.

#12 - 03/12/2016 07:31 PM - Luke Murphey

Trying to index without vectors

#13 - 03/12/2016 07:33 PM - Luke Murphey

Debugging:

- Try on Windows: still fails
- Try on older versions: different error
- Try without vectors: works (including no vectors on the content field)
- Index by verse to identify bad work:

#14 - 03/12/2016 09:16 PM - Luke Murphey

Indexing without vectors works.

#15 - 03/12/2016 11:06 PM - Luke Murphey

Ok, I have a fairly minimal repro:

```
python .\manage.py make_search_indexes -c -w "Cynegeticus"
python .\manage.py make_search_indexes  -w "the-deipnosophists-book-1"
```

After importing Cynegeticus, the index cannot be opened.

#16 - 03/12/2016 11:59 PM - Luke Murphey

Even smaller repro:

```
python .\manage.py make_search_indexes -c -w indica
```

#17 - 03/13/2016 06:38 PM - Luke Murphey

Opened ticket: https://bitbucket.org/mchaput/whoosh/issues/439/block-tag-error-vpst-generated-on-indexing

#18 - 03/13/2016 10:20 PM - Luke Murphey

Employed a workaround but this still didn't allow the entire index to be built.

Blew up after:

```
reader.contentsearch: Successfully indexed division, division="Book 12", work="Laws"
```

#19 - 03/13/2016 10:30 PM - Luke Murphey

I'm changing the analyzer such that it matches at least one character to see if that makes a difference.

#20 - 03/14/2016 01:18 AM - Luke Murphey

Nope, still fails.

#21 - 03/14/2016 05:04 PM - Luke Murphey

Trying without the analyzer.

#22 - 03/14/2016 05:08 PM - Luke Murphey

Interesting. Removing the analyzer causes the indexer to blow up much earlier (blows up on Argonautica).

#23 - 03/14/2016 07:20 PM - Luke Murphey

```
Logged from file contentsearch.py, line 250
Process SubWriterTask-416:
Traceback (most recent call last):
  File "C:\Program Files (x86)\Python2.7\lib\multiprocessing\process.py", line 258, in _bootstrap
    self.run()
  File "D:\Users\Luke\Workspace\TextCritical.com\src\whoosh\multiproc.py", line 129, in run
    runname, fieldnames, segment = finish_subsegment(writer, k)
  File "D:\Users\Luke\Workspace\TextCritical.com\src\whoosh\multiproc.py", line 49, in finish_subsegment
    runname = writer.pool.runs[0]
IndexError: list index out of range
```

#24 - 03/14/2016 08:03 PM - Luke Murphey

Indexing works provided I exclude:

- laws
- commentary-on-plato-protagoras-adam
- speeches-hyperides-english

#25 - 03/16/2016 07:02 PM - Luke Murphey

- Target version changed from 1.3 to 1.3.1

#26 - 03/17/2016 04:16 PM - Luke Murphey

- Status changed from New to In Progress

#27 - 03/18/2016 07:47 PM - Luke Murphey

- Status changed from In Progress to Closed

#28 - 03/18/2016 07:47 PM - Luke Murphey

- % Done changed from 0 to 100