TextCritical.net - Feature #1224

Morphology tool word search

03/02/2016 05:22 PM - Luke Murphey

Status:	Closed	**Start date:**	03/04/2016
Priority:	Normal	**Due date:**	
Assignee:	Luke Murphey	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	1.3		

Description

Add the ability to do a word search from the morphology tool that would provide:

- Count of this word in the work
- Count of this word in the section
- Pie chart of the forms (if searching for related forms)

Subtasks:

Task # 1235: Add Javascript charting library	**Closed**
Task # 1236: Add endpoint to return stats	**Closed**
Task # 1237: Add tabs to switch between results, stats and help	**Closed**
Feature # 1239: Chart of works matched	**Closed**
Feature # 1240: Show message when chart has no data	**Closed**

Related issues:

Related to TextCritical.net - Bug #1250: Word frequency chart is incorrect	**Closed**	**03/07/2016**

History

#1 - 03/02/2016 05:39 PM - Luke Murphey

It would be nice to be able to do this from the search tool too. That way one could get counts from several words in an OR search. Another option would just be adding a count of the number of word matches to the search page (right now it lists verse matches, not word matches).

#2 - 03/02/2016 08:00 PM - Luke Murphey

- Target version changed from 3.0 to 1.3

#3 - 03/03/2016 08:22 AM - Luke Murphey

If this is on the search page then this could include:

- The various forms if similar forms are searches for
- The count of forms by division if multiple are returned
- The count of each search term that matched

#4 - 03/03/2016 08:23 AM - Luke Murphey

Note sure how the highlights work and if the forms can be extracted.

#5 - 03/03/2016 05:44 PM - Luke Murphey

Doing some testing to see what the results object includes:

```
from reader.contentsearch import *
inx = WorkIndexer.get_index()
parser = QueryParser("content", inx.schema, termclass=GreekBetaCodeVariations)

searcher = inx.searcher()
```

```
def do_search(search_text):
    search_query = parser.parse(unicode(search_text))
    print search_query
    r = searcher.search_page(search_query, 1, 20, terms=True)
    print len(r.results)
    return r

r = do_search(u'NO/MOU')
```

This example provides the matched terms:

```
from reader.contentsearch import *
inx = WorkIndexer.get_index()
parser = QueryParser("content", inx.schema, termclass=GreekVariations)

searcher = inx.searcher()

def do_search(search_text):
    search_query = parser.parse(unicode(search_text))
    print search_query
    r = searcher.search_page(search_query, 1, 20, terms=True)
    print len(r.results)
    return r

r = do_search(u'NO/MOU')
r.results.matched_terms()
```

#6 - 03/03/2016 07:49 PM - Luke Murphey

Whoosh doesn't seem to store the number of matches in the matched terms. Perhaps a collector could do this:
http://whoosh.readthedocs.org/en/latest/api/collectors.html

#7 - 03/03/2016 07:58 PM - Luke Murphey

frequency() and doc_frequency() are helpful:

Number of examples of it:
searcher.frequency("content", u'εἰς')

Number of verses that include it:

searcher.doc_frequency("content", u'εἰς')

#8 - 03/03/2016 08:26 PM - Luke Murphey

```
s = searcher.postings("content", u'εἰς')
i = s.all_items()
i.next()
```

See http://whoosh.readthedocs.org/en/latest/recipes.html?highlight=doc_frequency

Might want to consider:

1. term vectors
2. Iterating through postings

#9 - 03/03/2016 09:25 PM - Luke Murphey

Might be able to:

1. Use searcher.postings() to match a term
2. Skip to the ID of the earliest verse in the document (presumes that documents are imported in order)
3. Iterate through results until the last Id within the document (or some limit)

#10 - 03/04/2016 01:11 AM - Luke Murphey

I think it turns out the Whoosh doesn't indicate the number of hits within a document. Instead it highlights them only when you provide the content.

#11 - 03/04/2016 01:40 AM - Luke Murphey

A raw SQL query works surprisingly well:

```
select * from reader_verse
inner join reader_division on reader_verse.division_id = reader_division.id
inner join reader_work on reader_work.id = reader_division.work_id
where
reader_work.title_slug = "new-testament"
AND reader_verse.content like "%καὶ%"
```

#12 - 03/04/2016 02:59 AM - Luke Murphey

r.results.termdocs indicates the terms that matched

#13 - 03/04/2016 03:04 AM - Luke Murphey

Things here don't make sense.

```
work:"New Testament" section:"Galatians" νόμον
```

This returns 25 matches (counted 32) and:
νομον: 16
νομος: 7
νομου: 8
νομῳ: 4

```
work:"New Testament" section:"Galatians" νόμος
```

This returns 7 matches and:
νομος: 7

```
work:"New Testament" section:"Galatians" νόμῳ
```

This returns 25 (counted 32 again) matches and:
νομον: 8
νομος: 7
νομου: 8
νομῳ: 8

#14 - 03/04/2016 03:16 AM - Luke Murphey

It turns out that only one form of νομος is being found: NOMOΣ

#15 - 03/06/2016 06:46 AM - Luke Murphey

Also getting too many responses. For example, the following returns 9 verses but 11 matches:

```
(work:new-testament)  xaris  (section:"Galatians 1" OR section:"Galatians 2" OR section:"Galatians 3" OR secti
on:"Galatians 4" OR section:"Galatians 5" OR section:"Galatians 6")
```

Four instances of χαρις are matched while I can only find 2. Matching on χαρις directly only returns two (either with or without diacritics). Also, no variations of χαρις are being found.

I'm wondering if variations include duplicates?

#16 - 03/06/2016 07:01 AM - Luke Murphey

These counts don't make complete sense.

The following indicates four matches each for νομον and εργον. However, there are 7 instances of νομου.

```
(work:new-testament) (νόμον εργόν) (section:"galatians 1" OR section:"Galatians 2" OR section:"Galatians 3" OR
 section:"Galatians 4" OR section:"Galatians 5" OR section:"Galatians 6")
```

#17 - 03/06/2016 07:05 AM - Luke Murphey

This seems to have something to do with variations.

The following indicates 10 matches in six verses (all of the search terms are exact matches):

```
(work:new-testament) (ἔργων) (section:"galatians 1" OR section:"Galatians 2" OR section:"Galatians 3" OR secti
on:"Galatians 4" OR section:"Galatians 5" OR section:"Galatians 6")
```

The following only shows 6 matches in six verses (using variations):

```
(work:new-testament) (εργόν) (section:"galatians 1" OR section:"Galatians 2" OR section:"Galatians 3" OR secti
on:"Galatians 4" OR section:"Galatians 5" OR section:"Galatians 6")
```

#18 - 03/06/2016 07:07 AM - Luke Murphey

Wow, this matches 16 in six verses:

```
(work:new-testament) (εργόν OR ἔργων) (section:"galatians 1" OR section:"Galatians 2" OR section:"Galatians 3"
 OR section:"Galatians 4" OR section:"Galatians 5" OR section:"Galatians 6")
```

#19 - 03/07/2016 05:28 AM - Luke Murphey

- Status changed from New to Closed

#20 - 03/09/2016 04:53 AM - Luke Murphey

- Related to Bug #1250: Word frequency chart is incorrect added