TextCritical.net - Feature #526

Add ability to search beta-code

02/16/2013 08:29 PM - Luke Murphey

Status:	Closed	**Start date:**	
Priority:	Normal	**Due date:**	
Assignee:	Luke Murphey	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	0.5		

Description

Provide users the ability to search using beta-code. In the best case scenario, I would like the results to highlight the beta-code sections.

This could be done by:

1. Converting the beta-code to a Greek character set in the search
2. Providing a mechanism for people to select Greek chars (like a drop-down Greek keyboard)
3. Indexing beta-code version of the content

The first and second options are nice because users can see the actual Greek content they are searching for and we don't have to index more content (which makes the indexes larger). Also, the search results will automatically highlight the Greek characters.

Related issues:

Blocks TextCritical.net - Feature #528: Search help page	**Closed**	**02/17/2013**

History

#1 - 02/16/2013 08:30 PM - Luke Murphey

Could also add a button to convert the beta-code to Unicode

#2 - 02/18/2013 05:24 AM - Luke Murphey

I think I may be able to make a stemming function that can help perform searches with beta-code.

- http://pythonhosted.org/Whoosh/stemming.html?highlight=stemming
- http://stackoverflow.com/questions/3688432/djangohaystackwhoosh-how-to-deal-with-language-inflection
- http://pythonhosted.org/Whoosh/api/analysis.html?highlight=stemming#whoosh.analysis.StemmingAnalyzer

#3 - 02/18/2013 05:33 AM - Luke Murphey

Actually, I may want to use variations: http://pythonhosted.org/Whoosh/api/query.html#whoosh.query.Variations

#4 - 02/18/2013 05:47 AM - Luke Murphey

- % Done changed from 0 to 30

I have variations partially working. The only problem is that the query parser is splitting the query based on non-word characters such as slashes. This means that a search for "NO/MOU" results in a search for "no" and "mou" as opposed to expanding into "νομου".

#5 - 02/18/2013 06:25 AM - Luke Murphey

I see information about how the Tokenizers split up the content for searching but I cannot find docs about how the searches are split up.

I think the SpaceSeparatedTokenizer is the correct one to use.

#6 - 02/18/2013 07:56 AM - Luke Murphey

This person seems to be running into the same problem:
http://stackoverflow.com/questions/14296792/haystack-whoosh-search-for-email-address-with-at-symbol/14482122#14482122

#7 - 02/18/2013 08:23 AM - Luke Murphey

Asked question on StackOverflow: http://stackoverflow.com/questions/14932013/include-slashes-and-parentheses-in-tokens

#8 - 02/18/2013 08:53 AM - Luke Murphey

It looks like the query parser uses the analyzer associated with the schema to parse the search. See QueryParser.__init__:

> :param schema: a :class:`whoosh.fields.Schema` object to use when
> parsing. The appropriate fields in the schema will be used to
> tokenize terms/phrases before they are turned into query objects.
> You can specify None for the schema to create a parser that does
> not analyze the text of the query, usually for testing purposes.

#9 - 02/18/2013 09:39 AM - Luke Murphey

Changing the regular expression that SimpleAnalyzer uses to "[\w/*()=\+|&']+(\.?\w+)*" does get the unit tests to pass. However, the searches still fail. I'm guessing that the analyzer gets persisted in the index file and thus the indexes have to be re-created.

#10 - 02/18/2013 09:57 AM - Luke Murphey

Re-indexing the database didn't seem to help. Doing a search for "pa/sxa" returns no results despite the fact that this is converted to "πασχα". Oddly enough, searching for "πασχα" does return results. These are actually different:

```
CF 80 E1 BD B1 CF 83 CF 87 CE B1
CF 80 CE AC    CF 83 CF 87 CE B1
```

Somehow, the unicode for the accent is getting saved differently.

#11 - 02/18/2013 10:00 AM - Luke Murphey

Normalizing the content before saving it in the index worked. The only problem is that highlighting doesn't happen.

#12 - 02/18/2013 10:02 AM - Luke Murphey

- % Done changed from 30 to 60

#13 - 02/18/2013 10:05 AM - Luke Murphey

The analyzer does seem to be pickled. I'll need to recreate the indexes from scratch.

#14 - 02/18/2013 08:51 PM - Luke Murphey

I rebuilt the indexes. The search works provided I normalize the unicode on the search term. The problem is that the terms are not being highlighted, either with a search for πασχα or pa/sxa.

#15 - 02/18/2013 08:52 PM - Luke Murphey

Hmm, not sure what is going on here. Highlighting is inconsistent. ἑορτὴ does get highlighted.

#16 - 02/18/2013 08:54 PM - Luke Murphey

Somehow the content is getting stored in a different format in the indexes. I normalized the content before I provide it to the highlights function and this fixed the problem.

#17 - 02/18/2013 08:55 PM - Luke Murphey

Nice, it looks like the beta-code causes the text to get highlighted correctly now. Yay

#18 - 02/18/2013 09:25 PM - Luke Murphey

- % Done changed from 60 to 80

I wonder: could I use the variations to search for all forms of a word? If so, then I could allow users to look up all possible forms of a given word.

#19 - 02/18/2013 10:25 PM - Luke Murphey

- Status changed from New to Closed

- % Done changed from 80 to 100