SerbianLanguageSupport

Serbian language uses both Cyrillic and Latin alphabet. Solr can search both at the same time (that is, search texts written in Cyrillic and Latin alphabet using queries written in Cyrillic or Latin alphabet). It is also possible to search with "bald" Latin alphabet (without diacritics).

To enable this feature, two Serbian-language filters can be used. Both are produced by `SerbianNormalizationFilterFactory` which accepts parameter `haircut`, with the value `regular` or `bald` (the default).

Filter choice

A corpus to be searched could contain Cyrillic, regular Latin and/or "bald" Latin; users likewise may enter their queries variously.

- If the search corpus contains "bald" Latin, simply use `SerbianNormalizationFilterFactory`.
- If the search corpus has only Cyrillic or regular Latin text, **and** the users can be expected to enter Cyrillic or regular Latin, use `SerbianNormalizationFilterFactory` with the parameter `haircut="regular"`.
- If the search corpus has only Cyrillic or regular Latin text, but users can be expected to search with "bald" Latin, there are two solutions:
    - To simply use `SerbianNormalizationFilterFactory` with slightly worse results.
    - To use two indices: one index should use `SerbianNormalizationFilterFactory` and the other should use `SerbianNormalizationFilterFactory` with `haircut="regular"` (you can use copyField directive to copy from one to the other). Then, if a user enters a query that contains a Cyrillic letter or any of ', ', 'š', 'ž' or ' (regexp: `[ašž]`), search only the regular index; otherwise (the query might be "bald"), search the "bald" index.

Background

Serbian language is specific in that it uses two alphabets, Cyrillic and Latin; while Cyrillic alphabet is considered the primary, Latin alphabet is also common. Texts might contain both alphabets, users might enter queries in both alphabets, so it is important to be able to search both at the same time.

Searching in both alphabets is easily achieved by converting Cyrillic to Latin, which is done by `SerbianNormalizationFilterFactory` with the parameter `haircut="regular"`. While the alphabets are not fully isomorphic, in practice this is not a problem for search, since converting from Cyrillic to Latin does not produce doublet words (there are practically no cases of two Serbian words written differently in Cyrillic that are written the same in Latin).

However, a potential problem for search is that a lot of Serbian speakers, especially in online usage, tend to use "bald" Latin alphabet, that is, Latin alphabet without diacritics. Searching in "bald" Latin is achieved by converting Cyrillic and regular Latin to "bald" Latin, which is done by `SerbianNormalizationFilterFactory` with the parameter `haircut="bald"` (or without a parameter). This does produce a number of doublet words that can cause problems with search, for example: "strašan" (horrifying) and "strasan" (passionate) or "teža" (gravity) and "teza" (thesis).

Hence, the filter choices above.

If the search corpus itself contains "bald" Latin, it already contains doublets, so converting the rest of the corpus to "bald" Latin is the best that could be done. Examples of such a corpus could be an Internet forum or an e-mail archive.

If the search corpus has only Cyrillic or regular Latin (examples of such a corpus could be a digitized book collection or a newspaper archive), choice of the filter is dependent on user behavior.

In a formal setting, where users can be educated to not enter queries with "bald" Latin, using the regular filter is the best choice.

In an informal setting (online), where users can be expected to enter queries with "bald" Latin, one possibility is to use the "bald" filter, introducing the doublets but being the simpler solution. The other possibility is the two-index solution: if a user enters a query that contains Cyrillic or a character with diacritics, we may assume that the user knows what is he doing, and use the regular index; if not, use the "bald" index.