

A List of Farsi Stopwords

Kazem Taghva, Russell Beckley, Mohammad Sadeh

ISRI Technical Report No. 2003-01
Information Science Research Institute
University of Nevada, Las Vegas
Las Vegas, NV 89154-4021
e-mail: taghva@isri.unlv.edu

In Information Retrieval (IR), words that do not carry any information are known as *stopwords*. Examples of stopwords in English are 'the', 'by', and 'this'. These words typically account for a major part of the index if they are not removed at indexing time. In TREC, the top 33 stopwords account for 30 percent of all the words[3].

Farsi like any other natural language contains many stopwords that account for a good percentage of all word occurrences. We are not aware of any study that has identified a stopwords list for Farsi. As one of the by products of our Farsi project[2], we identified a collection of Farsi documents for testing purposes. This collection consists of 1850 documents. These documents were collected within a six month period from many websites. Some of these websites originate in Iran and they typically represent electronic versions of popular Iranian newspapers and magazines. The rest of the articles are from websites originating in the U.S. or Europe. These documents mainly cover topics such as politics, economic issues associated with oil, social problems, and historical changes in the Middle East. We have used this collection to automatically identify a Farsi stopwords list based on the distribution of the words. Referring to well-known English stopword lists and to common sense, we manually edited the result to remove words that, though frequent in our collection, should not be considered stopwords in a general collection.

Further evaluation of this list revealed that it was incomplete. Among the Farsi stopwords are 12 verbs, each with a past tense and imperative form. Including all valid prefix

نمیکنیم	نمیکنید	نمیکنند	نمیکنم	نمیکنی
نمیکنندا	نکنیم	نکنید	نکنند	نکنم
نکنی	نکندا	بکنیم	بکنید	بکنم
بکنم	بکنی	بکنند	میکنیم	میکنید
میکنندا	میکنم	میکنی	میکنند	کنم
کنید	کنند	کنم	کنی	

Table 1: Variations of the verb کردن

Infinitive Form	Past Tense	Imperative
کردن	کرد	کن
بودن	بود	باش
شدن	شد	شو
داشتن	داشت	دار
خواستن	خواست	خواه
گفتن	گفت	گوی
دادن	داد	ده
گرفتن	گرفت	گیر
آمدن	آمد	آی
توانستن	توانست	توان
یافتن	یافت	یاب
آوردن	آورد	آور

Table 2: Farsi verbal stopwords

and suffix combinations, verbs have as many as 100 variations. Though a given verbal root may occur frequently, most of its variations occur infrequently. Therefore, a large number of variations of stopwords failed to make our list, though each variation, like its root, is a poor search term. For example, کردن has کرد and کن as its past tense and imperative forms. Table 1 shows all variations of the verb کردن.

Instead of listing all these variations, we list the past tense and imperative forms of the verb. All the variations can be mechanically built from these two forms. We employ our stemmer [1] to reduce these verbal variations to one of these two forms. Table 2 and Table 3 (p. 3) represent verbal and non-verbal stopwords in Farsi.

References

- [1] Kazem Taghva, Russell Beckley, and Mohammad Sadeh. A stemming algorithm for the farsi language. Technical Report 2003-02, Information Science Research Institute, University of Nevada, Las Vegas, August 2003.
- [2] Kazem Taghva, Ron Young, Jeffrey Coombs, Russell Beckley, Mohammad Sadeh, and Ray Pareda. Farsi searching and display technologies. In *Proceedings of SDIUT'03 the 2003 Symposium on Document Image Understanding Technology*, Greenbelt Maryland, April 2003.
- [3] I. Witten, A. Moffat, and T. Bell. *Managing Gigabytes: Compressing and indexing documents and images*. Morgan Kaufmann, 2nd edition, 1999.

سایر	چیز	مدت	همچنان	دیگران
بیرون	کنونی	کل	طی	جا
موارد	آنکه	کاملا	کامل	مثلا
بخشی	بطور	اکنون	امور	واقعی
حاضر	نوعی	عدم	چگونه	تحت
نگاه	خویش	کنار	مقابل	وضع
خیلی	تو	بنابراین	زمانی	درون
مختلف	اینجا	جز	خودش	بزرگ
قبل	آنجا	همچنین	نوع	توسط
ایشان	شاید	طور	اینها	جناح
ممکن	پیدا	مانند	طریق	جهت
بی	غیر	کسی	جای	کسانی
اخیر	وقتی	جدید	درباره	قابل
جریان	طرف	روی	بیش	چرا
چیزی	فقط	البتة	آنچه	زیر
زمینه	بخش	هنوز	برابر	چون
نشان	همان	استفاده	بدون	بین
اعلام	روز	عمل	بعد	بسیاری
تمام	امروز	بلکه	آنان	چند
دیگری	علیه	برخی	آیا	بیشتر
داده	حتی	انجام	گذشته	ویژه
حال	زمان	ولی	سوی	راه
همین	عنوان	یعنی	بسیار	تنها
اینکه	یکی	وی	پیش	هیچ
میان	چنین	پس	شما	وجود
نه	همه	اگر	چه	مورد
او	هر	باید	آنها	دیگر
اما	نیز	تا	من	ما
هم	یا	بر	خود	يك
برای	آن	با	این	را
از	که	به	در	و

Table 3: Non-verbal stopwords in Farsi