عنوان: **استفاده از لوسین برای انجام محاسبات آماری بر روی متون**
نویسنده: وحید نصیری
تاریخ: ۱۳۹۱/۰۹/۱۹ ۱۳:۳۳
آدرس: www.dotnettips.info
برچسب‌ها: Lucene.NET

احتمالا یک سری از کارهای اینفوگرافیک مانند tags cloud و words cloud را دیده‌اید. برای مثال در یک سخنرانی خاص، سخنران بیشتر از چه واژه‌هایی استفاده کرده است و سپس ترسیم درشت‌تر واژه‌هایی با تکرار بیشتر در یک تصویر نهایی. محاسبات آماری این نوع بررسی‌ها را توسط لوسین نیز می‌توان انجام داد که در ادامه به نحوه انجام آن خواهیم پرداخت.

**بررسی آماری واژه‌های بکار رفته در شاهنامه**

**مرحله اول: ایجاد ایندکس**

```
using System;
using System.Collections.Generic;
using System.IO;
using Lucene.Net.Analysis.Standard;
using Lucene.Net.Documents;
using Lucene.Net.Index;
using Lucene.Net.Store;

namespace ShaahnamehAnalysis
{
    public static class CreateIndex
    {
        static readonly Lucene.Net.Util.Version _version = Lucene.Net.Util.Version.LUCENE_CURRENT;

        static HashSet<string> getStopWords()
        {
            var result = new HashSet<string>();
            var stopWords = new[]
            {
                "به",
                "با",
                "از",
                "تا",
                "و",
                "است",
                "هست",
                "هستم",
                "هستیم",
                "هستید",
                "هستند",
                "نیست",
                "نیستم",
                "نیستیم",
                "نیستند",
                "اما",
                "یا",
                "این",
                "آن",
                "اینجا",
                "آنجا",
                "بود",
                "باد",
                "برای",
                "که",
                "دارم",
                "داری",
                "دارد",
                "داریم",
                "دارید",
                "دارند",
                "چند",
                "را",
                "ها",
                "های",
                "می",
                "هم",
                "در",
                "باشم",
```

```
"باشی",
"باشد",
"باشیم",
"باشید",
"باشند",
"اگر",
"مگر",
"بجز",
"جز",
"الا",
"اینکه",
"چرا",
"کی",
"چه",
"چطور",
"چی",
"چیست",
"آیا",
"چنین",
"اینچنین",
"نخست",
"اول",
"آخر",
"انتها",
"صد",
"هزار",
"میلیون",
"ملیون",
"میلیارد",
"ملیارد",
"یکهزار",
"تریلیون",
"تریلیارد",
"میان",
"بین",
"زیر",
"بیش",
"روی",
"ضمن",
"همانا",
"ای",
"بعد",
"پس",
"قبل",
"پیش",
"هیچ",
"همه",
"واما",
"شد",
"شده",
"شدم",
"شدی",
"شدیم",
"شدند",
"یک",
"یکی",
"نبود",
"میکند",
"میکنم",
"میکنیم",
"میکنید",
"میکنند",
"میکنی",
"طور",
"اینطور",
"آنطور",
"هر",
"حال",
"مثل",
"خواهم",
"خواهی",
"خواهد",
"خواهیم",
"خواهید",
"خواهند",
"داشته",
"داشت",
"داشتی",
"داشتم",
"داشتیم",
"داشتید",
```

```
                "داشتند",
                "آنکه",
                "مورد",
                "کنید",
                "کنم",
                "کنی",
                "کنند",
                "کنیم",
                "نکنم",
                "نکنی",
                "نکند",
                "نکنیم",
                "نکنید",
                "نکنند",
                "نکن",
                "بگو",
                "نگو",
                "مگو",
                "بنابراین",
                "بدین",
                "من",
                "تو",
                "او",
                "ما",
                "شما",
                "ایشان",
                "ی",
                "ـ",
                "هایی",
                "خیلی",
                "بسیار",
                "1",
                "بر",
                "1",
                "شود",
                "کرد",
                "کرده",
                "نیز",
                "خود",
                "شوند",
                "اند",
                "داد",
                "دهد",
                "گشت",
                "ز",
                "گفت",
                "آمد",
                "اندر",
                "چون",
                "بد",
                "چو",
                "همی",
                "پر",
                "سوی",
                "دو",
                "گر",
                "بی",
                "گرد",
                "زین",
                "کس",
                "زان",
                "جای",
                "آید"
            };

            foreach (var item in stopWords)
                result.Add(item);

            return result;
        }

        public static void CreateShaahnamehIndex(string file = "shaahnameh.txt")
        {
            var directory = FSDirectory.Open(new DirectoryInfo(Environment.CurrentDirectory +
"\\LuceneIndex"));
            var analyzer = new StandardAnalyzer(_version, getStopWords());
            using (var writer = new IndexWriter(directory, analyzer, create: true, mfl:
IndexWriter.MaxFieldLength.UNLIMITED))
            {
                var section = string.Empty;
                foreach (var line in File.ReadAllLines(file))
```

```
                {
                    int result;
                    if (int.TryParse(line, out result))
                    {
                        var postDocument = new Document();
                        postDocument.Add(new Field("Id", result.ToString(), Field.Store.YES,
Field.Index.NOT_ANALYZED));
                        postDocument.Add(new Field("Body", section, Field.Store.YES,
Field.Index.ANALYZED, Field.TermVector.WITH_POSITIONS_OFFSETS));
                        writer.AddDocument(postDocument);
                        section = string.Empty;
                    }
                    else
                        section += line;
                }

                writer.Optimize();
                writer.Commit();
                writer.Close();
                directory.Close();
            }
        }
    }
}
```

با ایجاد ایندکس‌های لوسین پیشتر در این سایت آشنا شده‌اید . روش کار نیز همانند سابق است. اطلاعات خود را، به هر فرمتی که تهیه شده باید تبدیل به اشیاء Document لوسین کرد. برای مثال در اینجا فقط یک فایل txt داریم که تشکیل شده است از تمام صفحات. به ازای هر صفحه، یک شیء Document تهیه و نوشته خواهد شد. همچنین در تهیه ایندکس از یک سری از واژه‌های بسیار متداول مانند «از»، «به»، «اندر»، (stopWords) صرفنظر شده است.

**مرحله دوم: ایجاد ابر واژه‌ها**

```
using System;
using System.Collections.Generic;
using System.Diagnostics;
using System.Linq;
using Lucene.Net.Index;
using Lucene.Net.Store;

namespace ShaahnamehAnalysis
{
    [DebuggerDisplay("{Frequency}, {Text}")]
    public class Tag
    {
        public string Text { set; get; }

        /// <summary>
        /// The frequency of a term is defined as the number of
        /// documents in which a specific term appears.
        /// </summary>
        public int Frequency { set; get; }
    }

    public static class WordsCloud
    {
        /// <summary>
        /// Create Words Cloud
        /// </summary>
        /// <param name="threshold">every term that appears in more than x Body</param>
        public static IList<Tag> Create(int threshold = 200)
        {
            var path = Environment.CurrentDirectory + "\\LuceneIndex";

            var results = new List<Tag>();
            var field = "Body";

            IndexReader indexReader = IndexReader.Open(FSDirectory.Open(path ), true);

            var termFrequency = indexReader.Terms();
            while (termFrequency.Next())
            {
                if (termFrequency.DocFreq() >= threshold && termFrequency.Term.Field == field)
                {
```

```
                    results.Add(new Tag { Text = termFrequency.Term.Text, Frequency =
termFrequency.DocFreq() });
                }
            }
            return results.OrderByDescending(x => x.Frequency).ToList();
        }
    }
}
```

پس از اینکه ایندکس لوسین تهیه شد، می‌توان به مداخل موجود در آن توسط متد indexReader.Terms دسترسی یافت.
نکته جالب آن فراهم بودن DocFreq هر واژه ایندکس شده است (فرکانس تکرار واژه؛ تعداد اشیاء Document ایی که واژه مورد نظر در آن‌ها تکرار شده است). برای مثال در اینجا اگر واژه‌ای 200 بار یا بیشتر در صفحات مختلف شاهنامه تکرار شده باشد، به عنوان یک واژه پر اهمیت انتخاب شده و به ابر واژه‌های نهایی اضافه می‌گردد.

**مرحله سوم: استفاده از نتایج**

```
using System;
using System.Diagnostics;
using System.IO;
using System.Linq;

namespace ShaahnamehAnalysis
{
    class Program
    {
        static void Main(string[] args)
        {
            CreateIndex.CreateShaahnamehIndex();
            var wordsCloudList = WordsCloud.Create();

            var data = wordsCloudList.Select(x => x.Text + ", " + x.Frequency)
                                     .Aggregate((s1, s2) => s1 + Environment.NewLine + s2);
            var output = "ShaahnamehAnalysis.txt";
            File.WriteAllText(output, data);
            Process.Start(output);
        }
    }
}
```

که نتیجه 15 مورد اول آن به صورت زیر است:

واژه | فرکانس
شاه, 1191
دل, 1088
سر, 1070
کار, 840
لشکر, 801
تخت, 755
روز, 745
ایران, 740
جهان, 724
مرد, 660
دست, 630
تاج, 623
نزدیک, 623
گیتی, 585
راه, 584

فایل‌های کامل این مثال را از اینجا می‌توانید دریافت کنید:

ShaahnamehAnalysis.zip