Automatic Topic Detection in Persian Blogs

Dr. Karine Megerdoomian and Dr. Ali Hadjarian
The MITRE Corporation

8 February 2011

MITRE Innovation Program

© 2011 The MITRE Corporation. All rights Reserved.

SMART POWER
A MITRE Corporation Initiative

Why Blogs?

- Bloggers provide a view into societies where polls may not be available
  - Audience analysis for strategic communication
  - Understanding culture and beliefs
  - Public opinion
- Cyber-communities emerge in the blogosphere
  - Analyze communities of interest
  - Identify influential bloggers in the network
  - See what issues and ideas bring people together
- Trends analysis
  - Study idea propagation by tracking issues through time
- But blogs are not always representative of the society
  - Depends on country's infrastructure, literacy rate, government filtering, popularity of blogs

Persian Blogosphere

- Estimates place Persian among the top ten languages of the global blog community
- Current blogs estimated at 700,000 with ~70,000 active blogs (updated at least once a week)
- The Persian blogosphere represents a wide spectrum of groups and opinions

Афсурдаги вакте ба сабк табдил мешавад
Ёдам меояд вакте Самарканд
хамаи сохтахои худ барои
мекардам, ба дустон ва хам
гирифтаам.

Challenges of Persian Blogs

- Persian *diglossia*
  - Two variants of the language, *literary* and *conversational*, are used side by side
  - The conversational variant affects the lexicon, word forms, and even syntax
  - Differences between major dialects (Farsi vs. Dari)
- Non-standard orthography
  - Spelling more directly represents the pronunciation of the words (e.g., *downloadz*, *dunno* in English)
  - Variation in spelling the same word across blogs or within a single blog
- Neologisms
  - Widespread use of borrowings, newly created words by bloggers, and new official terms by the Persian Language Academy
- Blog formats and encoding issues
  - The format and encoding used are not consistent across blogs or within a single blog

Automatic Topic Detection

- Develop automatic techniques for detecting topically related material in Persian language blogs.
- Combine language and text analytics with social network analysis to build a model of topic clusters
- Enhance system components:
  - Lexicon of blog terms
  - Morphological analyzer to process blog language
  - Topic classification system

Automatic Topic Detection

Persian Blog Post

با سلام

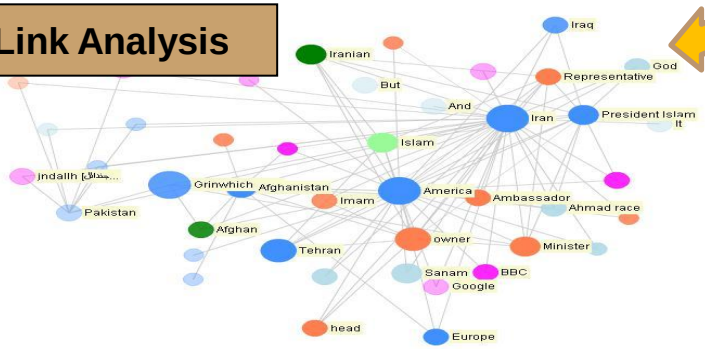
سالهاست در فضاي اينترنت زندگي مي كنم بخش مهمي از زندگي من با اينترنت آميخته شده است آن روزهاي اول كه وبلاگ نويسي را شروع كردم بيشتر مي خواستم از خاطرات دفاع مقدس مطلب بنويسم از دوستانم از بچه هاي سپاه از رشادتهايشان از

Morphological Analyzer

Lexicon

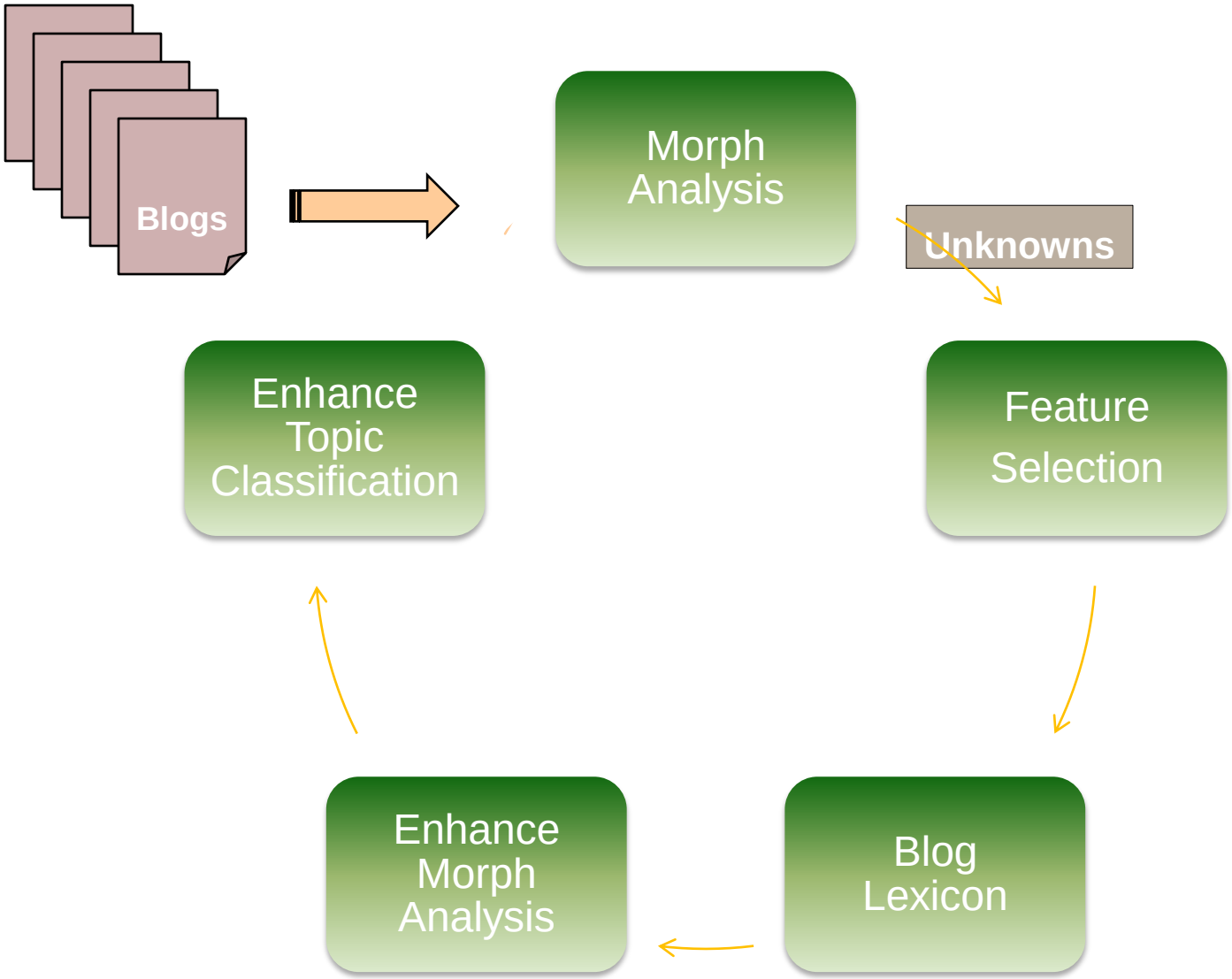
Topic Classification

Link Analysis



Technical Approach

- **Automatically extract neologisms**
  - **We employ morphological analysis in conjunction with a profile-based classification technique to extract a pertinent candidate list for identifying new word-level constructions in blogs.**
- **Build a Blog Lexicon**
  - **These neologisms are used to build a *blog lexicon* that is then integrated within the morphological analyzer.**
- **Enhance the topic classification system**
  - **We use the enhanced morphological analyzer and lexicon to extract a new set of features per topic category.**


Blogs
Morph Analysis
Unknowns
Feature Selection
Blog Lexicon
Enhance Morph Analysis
Enhance Topic Classification

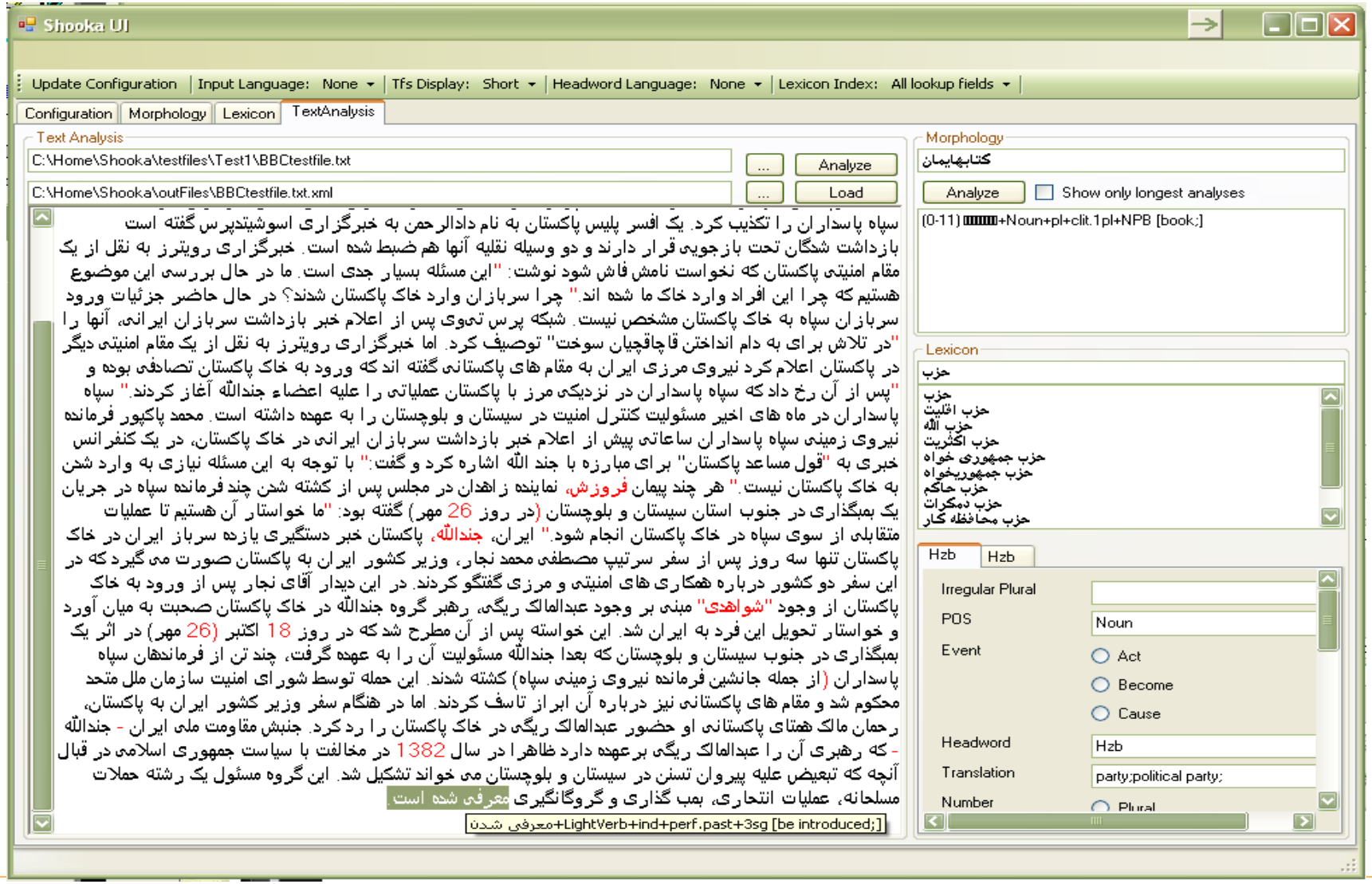
Blog Data

Topic	Number of blogs	Median size	Average number of words
Computers and Internet	497	14 kb	986
Cinema and theatre (sinama va ta'atr)	255	18 kb	1380
Political (siyasat-e-rooz)	500	22 kb	2171
Medical (pezeshki)	499	27 kb	2285
Sports (varzesh)	498	19 kb	1528

Linguistic Parsing

- Morphological Analyzer (Amtrup 2003)
  - Typed feature structures with unification
  - Full inflectional morphology, developed for MT application
  - Unanalyzed words are tagged as unknowns
  - 97% coverage and 93% accuracy on a 7 MB corpus collected from online news sources
- Lexicon used for lookup
  - About 40,000 entries in citation form and 5,000 proper names
  - Developed in 1999-2002 for coverage of online news articles
- Unanalyzed tokens
  - Conversational forms
  - Misspellings
  - Proper nouns
  - Specialized domain vocabulary
  - New words not in the lexicon

Linguistic Parsing



Information Gain

- Information Gain based feature selection
- Each attribute has a binary value which signifies the presence or absence of a given unknown term in the document

$$IG(D,t) = Entropy(D) - \frac{|D_t|}{|D|} Entropy(D_t) - \frac{|\overline{D_t}|}{|D|} Entropy(\overline{D_t})$$

- For a term to be selected, it not only needs to have a high IG, but it needs to be present in a higher proportion of positive examples than the negative ones
  - This prevents the selection of terms that while are good descriptors of the negative class and thus carry a high IG, are not necessarily pertinent to the positive class (i.e., the topic category under consideration). So IG of a term not meeting the above constraint is effectively set to zero.
- Extract neologisms that would have the most discriminatory power in distinguishing between topics

Weighted Unknowns for Computers

Transliteration		Weight	Translation
ویندوز	vyndvz	0.100033	Windows
دانلود	danlvd	0.080559	download
فایل	fayl	0.058319	file
کاربران	karbran	0.051595	users
جاوا	Java	0.048287	Java
کلیک	klyk	0.048180	click
یاهو	yahv	0.044999	Yahoo
نوکیا	nvkya	0.044807	Nokia
فلش	flG	0.042718	Flash
مرورگر	mrvrgr	0.041374	browser
هک	hk	0.041074	hack
مسنجر	msnJr	0.040853	Messenger
چت	Ct	0.039987	chat
پسورد	psvrd	0.039213	password
کد	kd	0.035936	code

Blog Lexicon

1. Automatically Extract Neologisms

- Collected a corpus of Persian language blogs in 5 major topics (politics, cinema, computers, medicine, sports)
- Used a morphological analyzer to tag unknowns
- Generated a weight-ordered list of unknown words for each topic category using information gain
- Words with English translations extracted from Wikipedia (Burger 2009)

Blog Lexicon

2. Manually collected terms

- Collected words from existing glossaries
  - Nuclear terms
  - Slang terms
  - Newly created technical and scientific terms from the Persian Language Academy
- Performed corpus analysis
  - Studied political and computer blogs to detect new usages that were not found in traditional lexicons using concordance tools

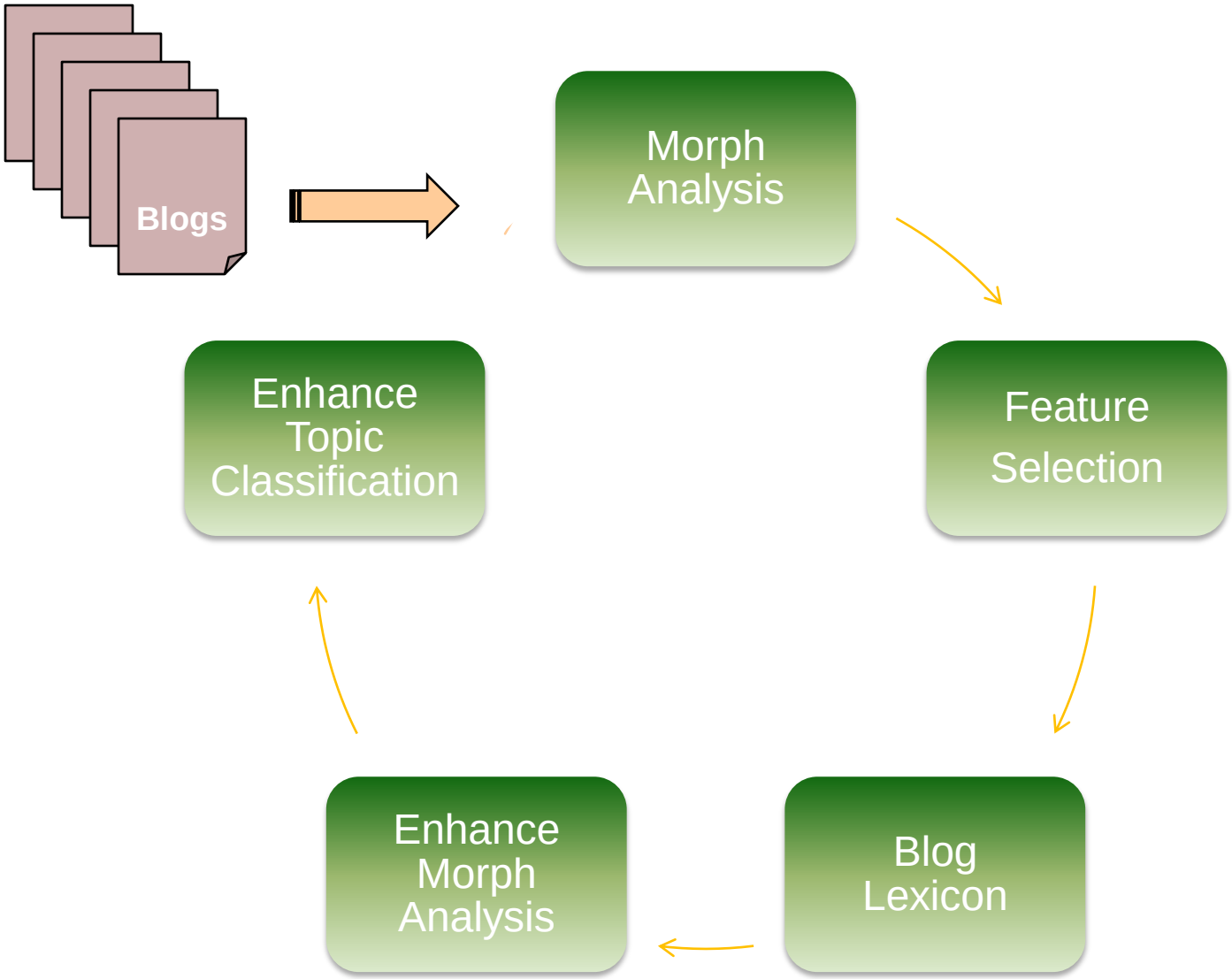
Blog Lexicon

- **Contains about 3,000 terms in 10 distinct categories**
- **Includes the Persian script, transcription, English translation and linguistic category for each term (e.g., English loan, abbreviation)**

Topic	Number	Topic	Number
Technical	1840	Science	89
General	332	Medical	58
Politics	257	Film	30
Nuclear	150	Slang	31
Military	116	Religion	26

Blog Lexicon Examples

Persian	Transcription	Formation	Translation	Domain
مرورگر	*morurgar*	browse + -er	Browser	Computer
دگرباش	*degarbâsh*	other + being	Queer	Politics/Society
بالگرد	*bâlgard*	wing + turn	Helicopter	Military
نمایشگر	*namâyeshgar*	show + -er	Monitor	Computer
مونیتور	*monitor*	(English loan)	Monitor	Computer
آوردنمایی	*âvardnamâyi*	input + showing	Battlefield Visualization	Military
فتنه گر	*fetnegar*	sedition + -er	Seditious	Politics
وبلاگستان	*veblâgestân*	weblog + -stan	Blogosphere	Computer
پیشدستانه	*spishdastâne*	pre + handed + ly	Preemptive	Military
چتیدن	*chatidan*	chat + (verb)	To chat	Computer
اس ام اس زدن	*es-em-es zadan*	SMS + hit	To send a text message	Computer
ذوالحقوق	*zolhoquq*	(Arab loan)	Righteous	Politics


Blogs
Morph Analysis
Feature Selection
Blog Lexicon
Enhance Morph Analysis
Enhance Topic Classification

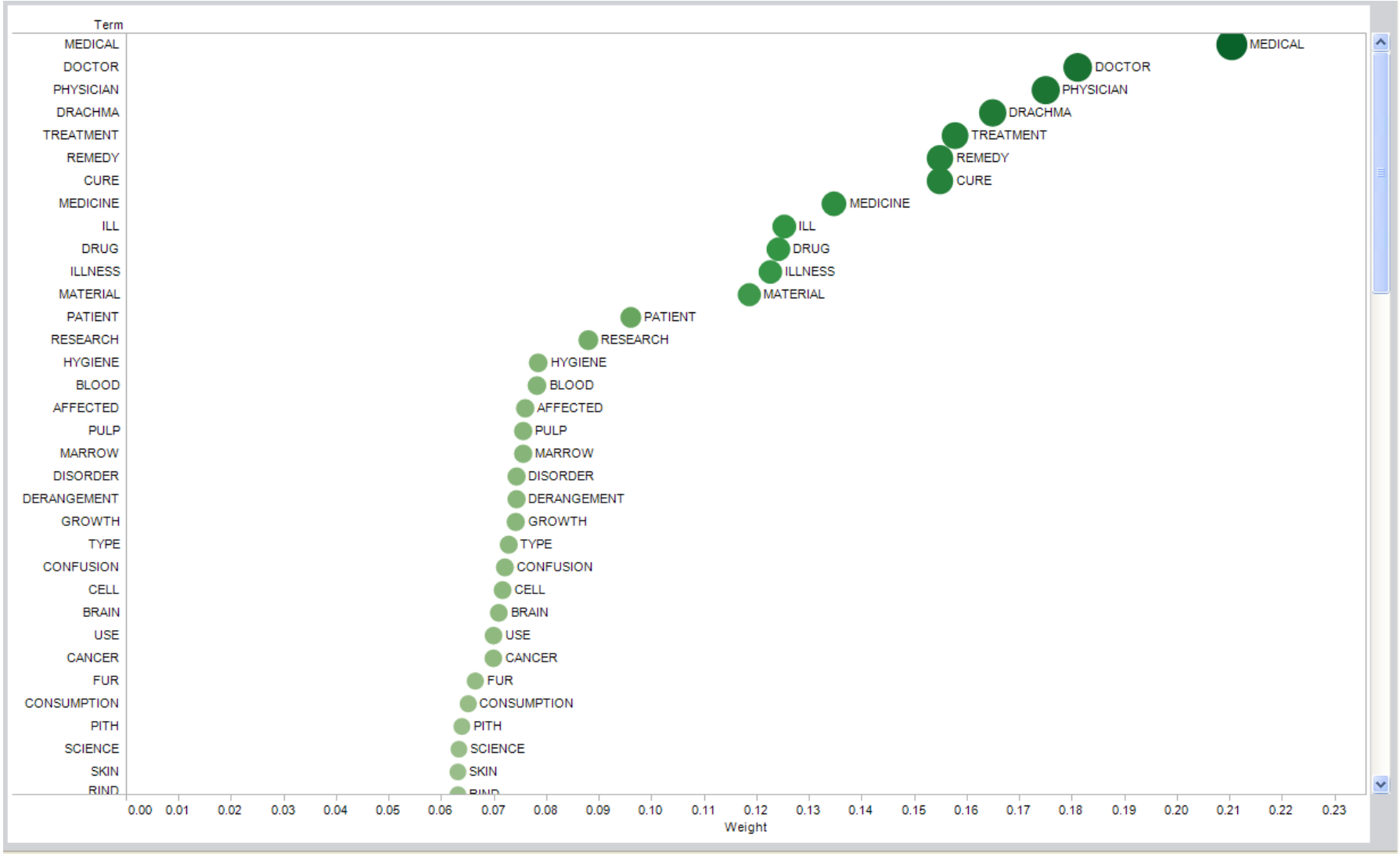
Weighted Terms per Category

1	Medical	11	Material
2	Doctor	12	Patient
3	Physician	13	Research
4	Treatment	14	Hygiene
5	Remedy	15	Blood
6	Cure	16	Affected
7	Medicine	17	Marrow
8	Ill	18	Pulp
9	Drug	19	Disorder
10	Illness	20	Derangement

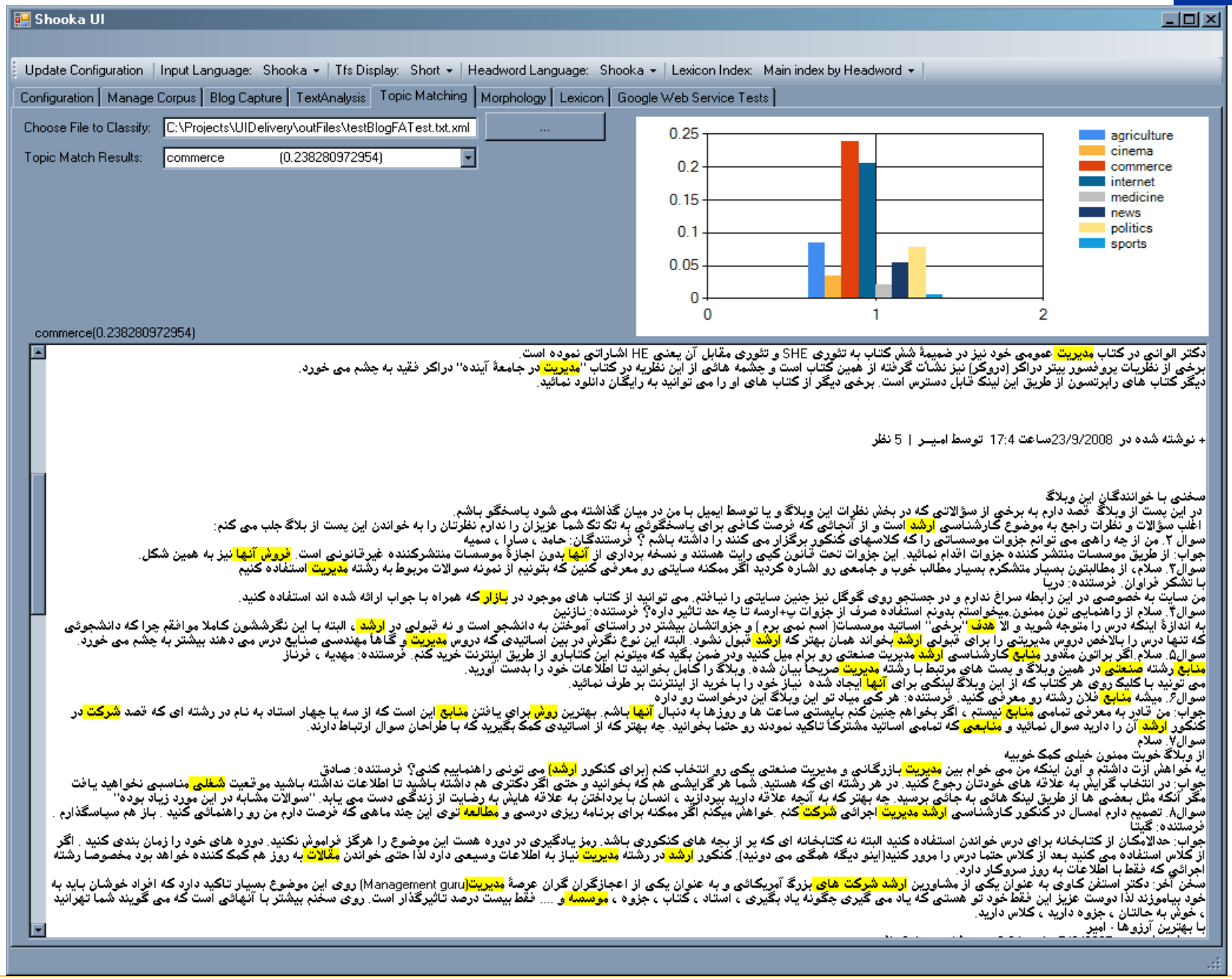
Top 20 terms for Medicine

- **Run blogs through enhanced morphological analyzer to obtain lemma or citation forms for each topic category**
- **Profile-based classifier trained on the terms (words) in the documents for each category**
- **Weighted terms used to identify the most significant key terms per topic**
- **Category of a blog post can then be determined by the similarity of its vector representation to the prototype vector for each topic**

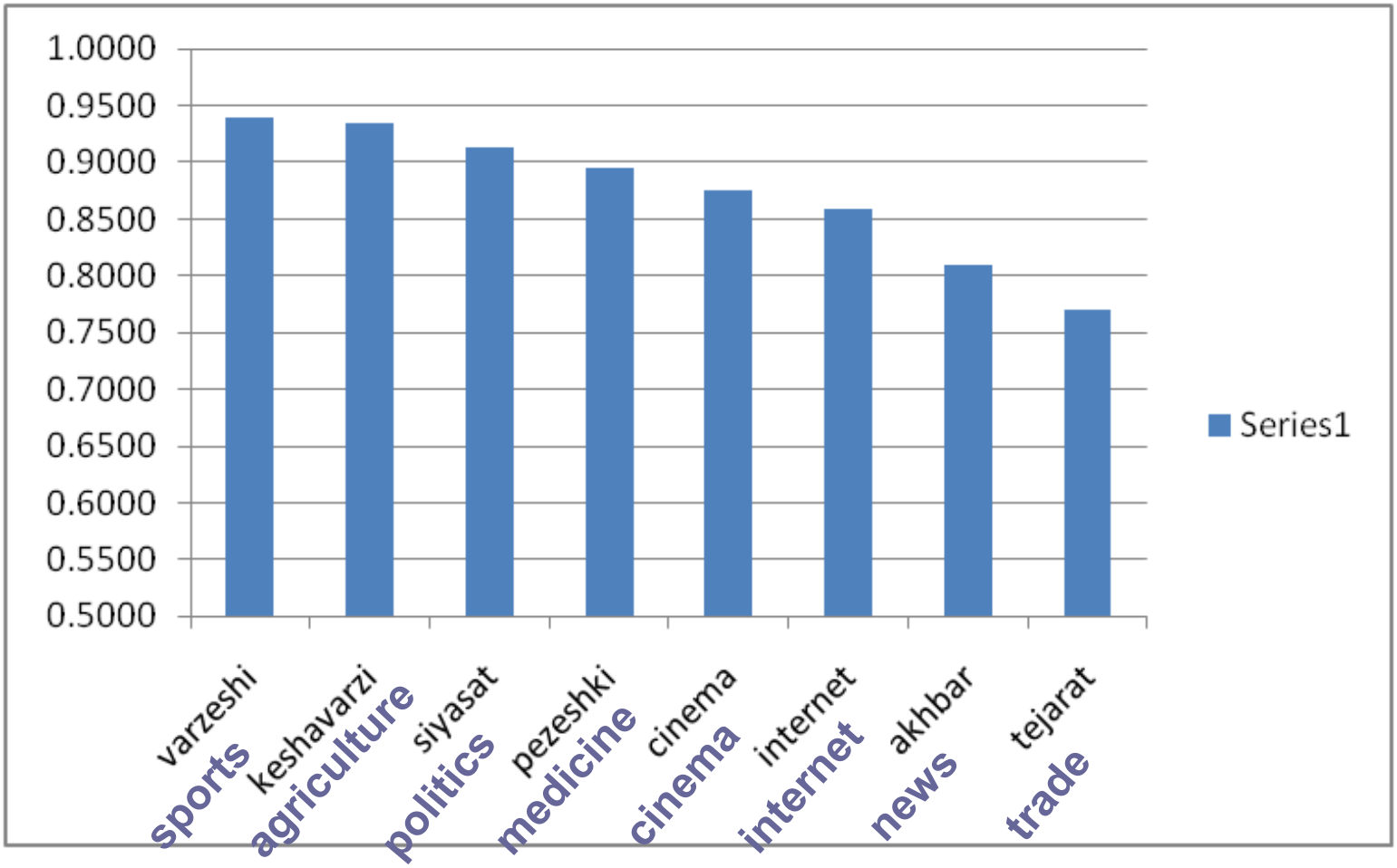
Features for the Medical Domain



MITRE


Shooka UI
Update Configuration
Input Language: Shooka
Tfs Display: Short
Headword Language: Shooka
Lexicon Index: Main index by Headword
Configuration
Manage Corpus
Blog Capture
TextAnalysis
Topic Matching
Morphology
Lexicon
Google Web Service Tests
Choose File to Classify:
C:\Projects\UIDelivery\outFiles\testBlogFATest.txt.xml
Topic Match Results:
commerce (0.238280972954)
0.25
0.2
0.15
0.1
0.05
0
0
1
2
agriculture
cinema
commerce
internet
medicine
news
politics
sports
commerce(0.238280972954)
دکتر الوانی در کتاب مدیریت عمومی خود نیز در ضمیمهٔ شش کتاب به تئوری SHE و تئوری مقابل آن یعنی HE اشاراتی نموده است.
برخی از نظریات پروفسور پیتر دراکر [دروکر] نیز نشأت گرفته از همین کتاب است و چشمه هائی از این نظریه در کتاب "مدیریت در جامعهٔ آینده" دراکر فقید به چشم می خورد.
دیگر کتاب های رابرتسون از طریق این لینک قابل دسترس است. برخی دیگر از کتاب های او را می توانید به رایگان دانلود نمائید.
+ نوشته شده در 23/9/2008ساعت 17:4 توسط امیـــر | 5 نظر
سخنی با خوانندگان این وبلاگ
در این پست از وبلاگ قصد دارم به برخی از سؤالاتی که در بخش نظرات این وبلاگ و یا توسط ایمیل با من در میان گذاشته می شود پاسخگو باشم.
اغلب سؤالات و نظرات راجع به موضوع کارشناسی ارشد است و از آنجائی که فرصت کافی برای پاسخگوئی به تک تک شما عزیزان را ندارم نظرتان را به خواندن این پست از بلاگ جلب می کنم:
سوال ۲. من از چه راهی می توانم جزوات موسساتی را که کلاسهای کنکور برگزار می کنند را داشته باشم ؟ فرستندگان: حامد ، سارا ، سمیه
جواب: از طریق موسسات منتشر کننده جزوات اقدام نمائید. این جزوات تحت قانون کپی رایت هستند و نسخه برداری از آنها بدون اجازهٔ موسسات منتشرکننده غیرقانونی است. فروش آنها نیز به همین شکل.
سوال۳. سلام، از مطالبتون بسیار متشکرم بسیار مطالب خوب و جامعی رو اشاره کردید اگر ممکنه سایتی رو معرفی کنین که بتونیم از نمونه سوالات مربوط به رشته مدیریت استفاده کنیم
با تشکر فراوان. فرستنده: دریا
من سایت به خصوصی در این رابطه سراغ ندارم و در جستجو روی گوگل نیز چنین سایتی را نیافتم. می توانید از کتاب های موجود در بازار که همراه با جواب ارائه شده اند استفاده کنید.
سوال۴. سلام از راهنمایی تون ممنون.میخواستم بدونم استفاده صرف از جزوات پ+ارسه تا چه حد تاثیر داره؟ فرستنده: نازنین
به اندازهٔ اینکه درس را متوجه شوید و الا هدف "برخی" اساتید موسسات(اسم نمی برم) و جزواتشان بیشتر در راستای آموختن به دانشجو است و نه قبولی در ارشد ، البته با این نگرششون کاملا موافقم چرا که دانشجوئی که تنها درس را بالاخص دروس مدیریتی را برای قبولی ارشد بخواند همان بهتر که ارشد قبول نشود. البته این نوع نگرش در بین اساتیدی که دروس مدیریت و گاهاً مهندسی صنایع درس می دهند بیشتر به چشم می خورد.
سوال۵. سلام.اگر براتون مقدور منابع کارشناسی ارشد مدیریت صنعتی رو برام میل کنید ودر ضمن بگید که میتونم این کتابارو از طریق اینترنت خرید کنم. فرستنده: مهدیه ، فرناز
منابع رشته صنعتی در همین وبلاگ و پست های مرتبط با رشته مدیریت صریحاً بیان شده. وبلاگ را کامل بخوانید تا اطلاعات خود را بدست آورید.
می تونید با کلیک روی هر کتاب که از این وبلاگ لینکی برای آنها ایجاد شده نیاز خود را با خرید از اینترنت بر طرف نمائید.
سوال۶. میشه منابع فلان رشته رو معرفی کنید. فرستنده: هر کی میاد تو این وبلاگ این درخواست رو داره
جواب: من قادر به معرفی تمامی منابع نیستم ، اگر بخواهم چنین کنم بایستی ساعت ها و روزها به دنبال آنها باشم. بهترین روش برای یافتن منابع این است که از سه یا چهار استاد به نام در رشته ای که قصد شرکت در کنکور ارشد آن را دارید سوال نمائید و منابعی که تمامی اساتید مشترکاً تاکید نمودند رو حتما بخوانید. چه بهتر که از اساتیدی کمک بگیرید که با طراحان سوال ارتباط دارند.
سوال۷. سلام
از وبلاگ خوبت ممنون خیلی کمک خوبیه
یه خواهش ازت داشتم و اون اینکه من می خوام بین مدیریت بازرگانی و مدیریت صنعتی یکی رو انتخاب کنم [برای کنکور ارشد] می تونی راهنماییم کنی؟ فرستنده: صادق
جواب: در انتخاب گرایش به علاقه های خودتان رجوع کنید. در هر رشته ای که هستید. شما هر گرایشی هم که بخوانید و حتی اگر دکتری هم داشته باشید تا اطلاعات نداشته باشید موقعیت شغلی مناسبی نخواهید یافت
مگر آنکه مثل بعضی ها از طریق لینک هائی به جائی برسید. چه بهتر که به آنچه علاقه دارید بپردازید ، انسان با پرداختن به علاقه هایش به رضایت از زندگی دست می یابد. "سوالات مشابه در این مورد زیاد بوده"
سوال۸. تصمیم دارم امسال در کنکور کارشناسی ارشد مدیریت اجرائی شرکت کنم .خواهش میکنم اگر ممکنه برای برنامه ریزی درسی و مطالعه توی این چند ماهی که فرصت دارم من رو راهنمائی کنید . باز هم سپاسگذارم .
فرستنده: گیتا
جواب: حدالامکان از کتابخانه برای درس خواندن استفاده کنید البته نه کتابخانه ای که پر از بچه های کنکوری باشد. رمز یادگیری در دوره هست این موضوع را هرگز فراموش نکنید. دوره های خود را زمان بندی کنید . اگر
از کلاس استفاده می کنید بعد از کلاس حتما درس را مرور کنید(اینو دیگه همگی می دونید). کنکور ارشد در رشته مدیریت نیاز به اطلاعات وسیعی دارد لذا حتی خواندن مقالات به روز هم کمک کننده خواهد بود مخصوصا رشته
اجرائی که فقط با اطلاعات به روز سروکار دارد.
سخن آخر: دکتر استفن کاوی به عنوان یکی از مشاورین ارشد شرکت های بزرگ آمریکائی و به عنوان یکی از اعجازگران گران عرصهٔ مدیریت(Management guru) روی این موضوع بسیار تاکید دارد که افراد خوشان باید به
خود بیاموزند لذا دوست عزیز این فقط خود تو هستی که یاد می گیری چگونه یاد بگیری ، استاد ، کتاب ، جزوه ، موسسه و .... فقط بیست درصد تاثیرگذار است. روی سخنم بیشتر با آنهائی است که می گویند شما تهرانید
، خوش به حالتان ، جزوه دارید ، کلاس دارید.
با بهترین آرزوها - امیر

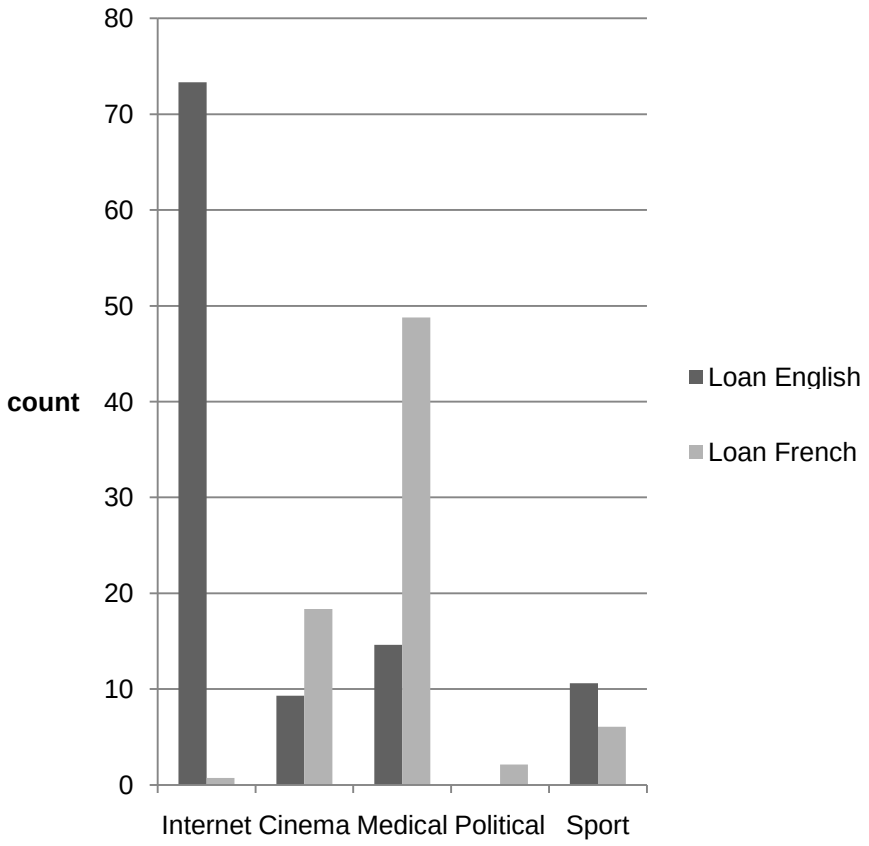
Topic Classification

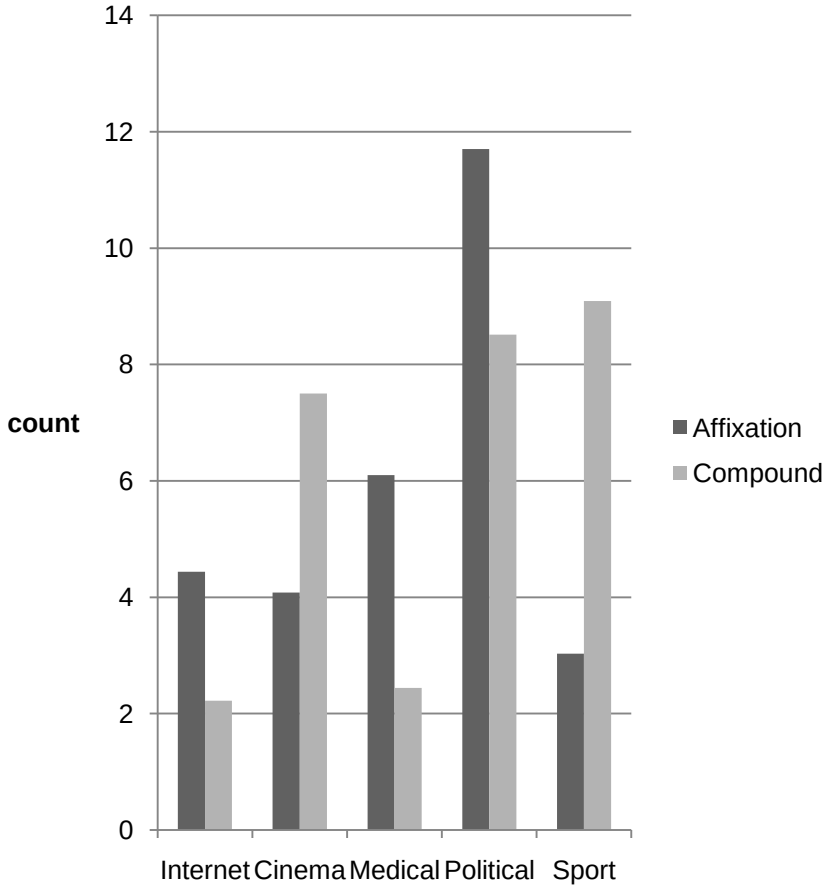


Language Correlations

- Weighted features selected by classifier correlates with topic
- Type of neologism used correlates with topic
  - Internet and computer blogs have lots of English loans
  - Political blogs create words based on Persian rules
- Type of neologism used correlates with author
  - Younger bloggers use English loans rather than use the official technical terms
  - It seems that younger bloggers use conversational language more freely

Language Correlations



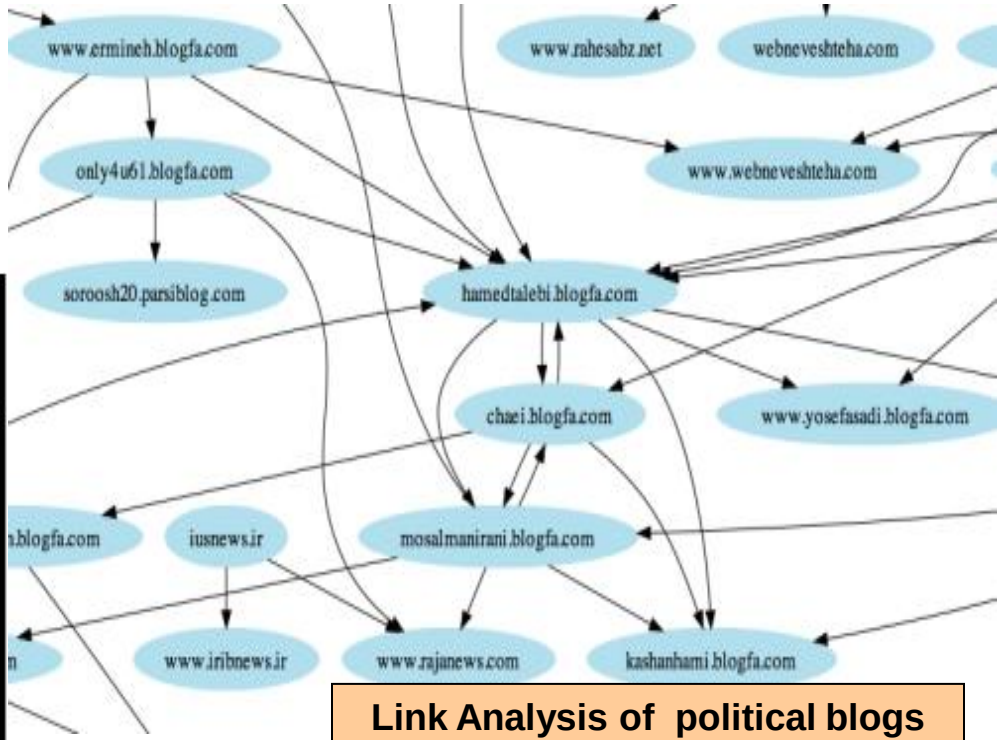


Link Structure Analysis

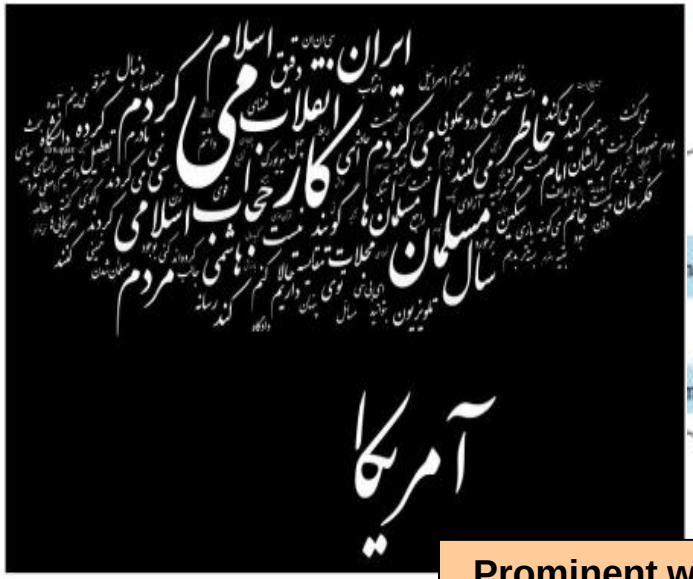
- Background
  - Investigating how blogs are connected through hyperlinks
  - Blogs that link to each other share issues and ideologies

- Automatically extracted links from blogs
- Combined with the results of classification to identify virtual clusters based on topic

Politics Topic Cluster



Link Analysis of political blogs



Prominent words for political cluster

America, work, muslim, revolution, year, justice, sedition, sacrifice, Iran, hijab, people, soft war, trial, Islamic, magazines, Hashemi, comparison, Imam…

Conclusion

- Built a Blog Lexicon
- Extended a morphological analyzer to process word forms encountered in blog language
- Developed a profile-based topic classifier for Persian blogs
- We combine the topic classifier with the results of link structure analysis to detect virtual topic clusters
- Future Work
  - Integrate results of sentiment analysis to identify bloggers' opinion(s) within topic clusters
  - Temporal analysis of topics in blog clusters