웹 로그인 아이디와 사용자 성별 및 연령대 관계 분석과 예측 모형
- 구조적 분석

2015. 11. 13.
박희웅
서울대학교 산업공학과

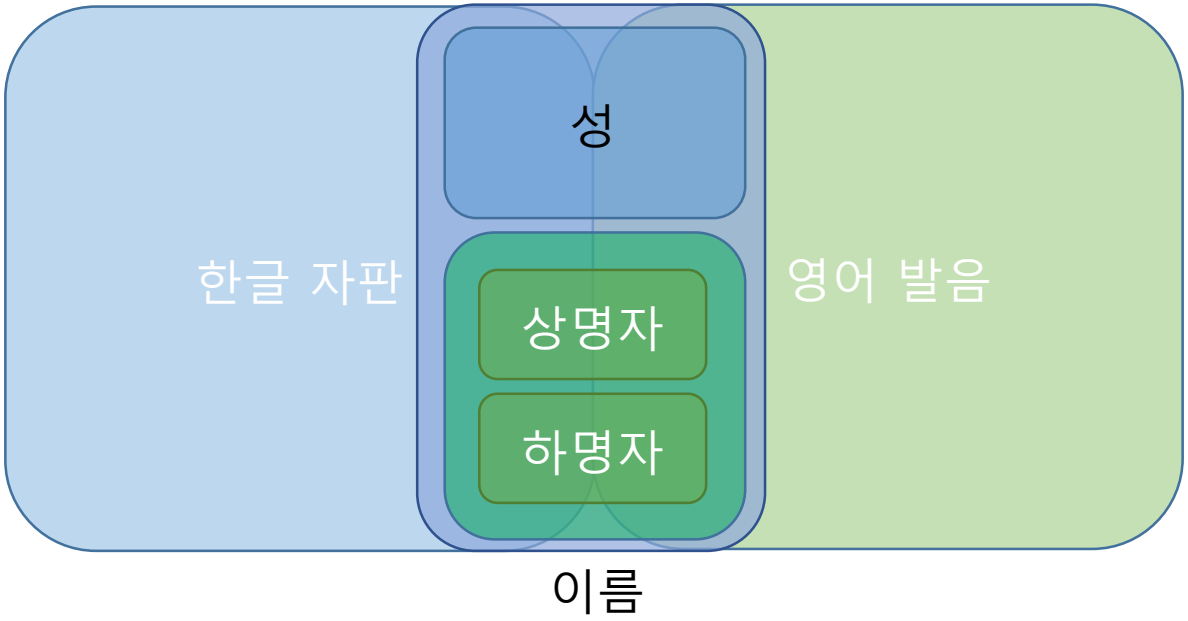
판별력 분석

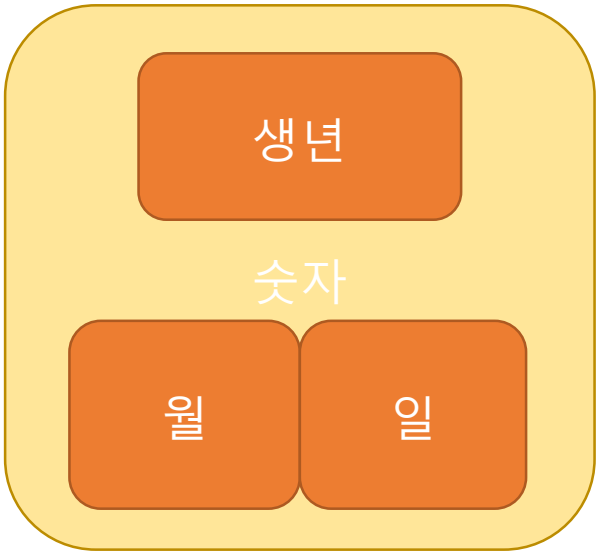
- 성별 판별력이 높은 형태소

morph	P(m)	P(m|F)	P(m|M)	성향
dms	0.653%	1.073%	0.395%	F
love	0.980%	1.392%	0.726%	F
gml	0.677%	1.042%	0.452%	F
al	0.453%	0.743%	0.274%	F
a	1.685%	2.063%	1.451%	F
hye	0.130%	0.289%	0.032%	F
rhdwn	0.087%	0.217%	0.006%	F
gp	0.197%	0.382%	0.083%	F
wl	0.598%	0.856%	0.439%	F
alswl	0.087%	0.206%	0.013%	F
eun	0.248%	0.423%	0.140%	F
fls	0.138%	0.279%	0.051%	F
girl	0.067%	0.165%	0.006%	F
thwjd	0.051%	0.134%	0.000%	F
ql	0.138%	0.268%	0.057%	F
thdud	0.047%	0.124%	0.000%	F
dbfla	0.047%	0.124%	0.000%	F
dbwls	0.091%	0.196%	0.026%	F

구조적 분석

- 아이디의 성분 별 구조화
  - 아이디를 형태소 단위로 분해한 뒤 각 요소에 의미적인 태그를 달아 분류 및 분석을 용이하게 할 수 있지 않을까?
  - 예: hee188 → hee + 188 → 하명자 + 숫자
  - 이름자는 성, 이름 그리고 이름은 다시 상명자와 하명자로 구성
  - 숫자의 경우, 생년, 월, 일, 이외 등으로 나눠볼 수 있음
  - 한글 자판을 그대로 친 것인지, 영어 발음으로 쓴 것인지 나눠볼 수 있음





구조적 분석

- 이용 데이터

한글 통계

상명자 "희"	남자 전체 이름중 23번째 (1.3064%)로 많이 쓰이며 상명자에서 33번째 (0.9624%) 하명자에서 14번째(1.6516%)로 많이 쓰입니다. 여자 전체 이름중 5번째 (4.0533%)로 많이 쓰이며 상명자에서 21번째 (1.6182%) 하명자에서 2번째(6.4914%)로 많이 쓰입니다.
하명자 "웅"	남자 전체 이름중 83번째 (0.275%)로 많이 쓰이며 상명자에서 137번째 (0.0756%) 하명자에서 51번째(0.4747%)로 많이 쓰입니다. 여자 전체 이름중 211번째 (0.006%)로 많이 쓰이며 상명자에서 229번째 (0.0024%) 하명자에서 174번째(0.0096%)로 많이 쓰입니다.

대한민국의 인구순 성씨 목록

위키백과, 우리 모두의 백과사전.

대한민국의 인구순 성씨 목록은 2000년 통계청 조사 결과를 바탕으로 작성된 인구순 성씨 목록이다. **굵은 글씨**는 동음이의 성이다.

목차 [숨기기]

1 대한민국의 성씨 인구 순위
2 대한민국 본관별 인구 100대 성씨
3 주석
4 바깥 고리

대한민국의 성씨 인구 순위 [편집]

2000년 조사				1985년 조사			순위 변화
순위	성씨	가구수	인구수	순위	가구수	인구수	
1	김(金)	3,102,537	9,925,949	1	2,080,768	8,785,341	-
2	이(李)	2,113,007	6,794,637	2	1,418,948	5,985,056	-
3	박(朴)	1,215,918	3,895,121	3	815,237	3,435,858	-
4	최(崔)	676,773	2,169,704	4	454,697	1,913,329	-
5	정(鄭)	626,265	2,010,117	5	422,246	1,780,734	-
6	강(姜)	325,288	1,044,386	6	227,097	958,181	-
7	조(趙)	306,022	984,913	7	207,896	877,058	-
8	윤(尹)	294,708	948,600	8	198,252	834,121	-
9	장(張)	287,195	919,339	9	192,842	810,235	-
10	임(林)	237,145	762,767	10	159,376	672,782	-

언어변환기 한글 이름 로마자 표기

박희웅 실행

한글 이름 로마자 표기

박희웅에 대해 현행 로마자 표기법(문화관광부 고시 2000-8호)에 의한 한글 이름 변환 결과 입니다.
이용자 편의를 위해 성은 일반적으로 통용되는 표기를 추천하며, 이름만 현행 로마자 표기법에 따라 표기합니다.

성(姓) : **박** Bak Huiung Park Huiung

사용 빈도 순

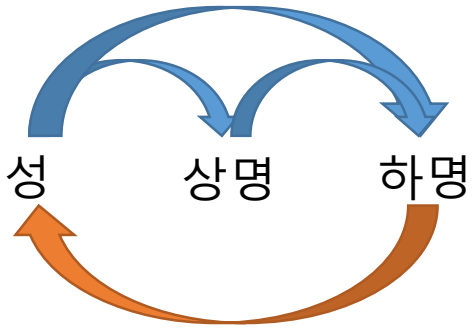
박희웅에 대해 일반적으로 많이 사용되는 한글 이름 로마자 표기 변환 결과입니다.
웹에서 사용되는 로마자 표기를 수집하여, 이 중 가장 많이 사용되는 결과를 추천합니다.

성(姓) : **박**

순위	한글 이름 로마자 표기	사용빈도
1	Park Heewoong	
2	Park Heeung	
3	Park Heewong	
4	Park Heewung	
5	Bak Heewoong	
6	Park Huiwoong	
7	Bak Heeung	
8	Park Huiung	
9	Bak Heewong	
10	Park Huiwong	

구조적 분석

- 은닉 마코프 모형 (HMM)



parkheewoong
parkwoong
heewoongpark

transition matrix 초기값

	한글 성	한글 상명	한글 하명	영문 성	영문 상명	영문 하명	그 외	끝
한글 성	0	1	1	0	0	0	1	1
한글 상명	0	0	1	0	0	0	1	1
한글 하명	1	0	0	0	0	0	1	1
영문 성	0	0	0	0	1	1	1	1
영문 상명	0	0	0	0	0	1	1	1
영문 하명	0	0	0	1	0	0	1	1
그 외	1	1	1	1	1	1	1	1
끝	0	0	0	0	0	0	0	1

Start probability 초기값

한글 성	한글 상명	한글 하명	영문 성	영문 상명	영문 하명	그 외	끝
1	1	1	1	1	1	3	0

- Emission probability
  - ✓ 이름자의 경우, 통계 정보 활용하여 확률 값 고정
  - ✓ 그 외 state emission probability는 초기값으로 모든 형태소에 동일한 확률 부여하고, 이후에 업데이트

구조적 분석

- HMM
  - 140만개 아이디로 학습

업데이트된 transition probability

	한글 성	한글 상명	한글 하명	영문 성	영문 상명	영문 하명	그 외	끝
한글 성	0	0.4863	0.0574	0	0	0	0.3636	0.0927
한글 상명	0	0	0.6780	0	0	0	0.3016	0.0203
한글 하명	0.0413	0	0	0	0	0	0.1389	0.8199
영문 성	0	0	0	0	0.2166	0.0808	0.3786	0.3240
영문 상명	0	0	0	0	0	0.4715	0.4949	0.0336
영문 하명	0	0	0	0.0293	0	0	0.2318	0.7389
그 외	0.0025	0.0090	0.0173	0.0158	0.0316	0.0724	0.4182	0.4332
끝	0	0	0	0	0	0	0	1

업데이트된 start probability

한글 성	한글 상명	한글 하명	영문 성	영문 상명	영문 하명	그 외	끝
4.963%	5.972%	0.061%	5.668%	16.276%	0.887%	66.172%	0

구조적 분석

- HMM (성, 이름만 분리)

모형 초기값

	한글 성	한글 이름	영문 성	영문 이름	그 외	끝
한글 성	0	1	0	0	1	1
한글 이름	1	1	0	0	1	1
영문 성	0	0	0	1	1	1
영문 상명	0	0	1	1	1	1
그 외	1	1	1	1	1	1
끝	0	0	0	0	0	1

한글 성	한글 이름	영문 성	영문 이름	그 외	끝
1	1	1	1	3	0

업데이트된 확률 값

	한글 성	한글 이름	영문 성	영문 이름	그 외	끝
한글 성	0	0.5757	0	0	0.4078	0.0165
한글 이름	0.0091	0.3606	0	0	0.1902	0.4401
영문 성	0	0	0	0.2876	0.4206	0.2918
영문 이름	0	0	0.0125	0.2668	0.3258	0.3949
그 외	0.0016	0.0263	0.0161	0.0988	0.4231	0.4341
끝	0	0	0	0	0	1

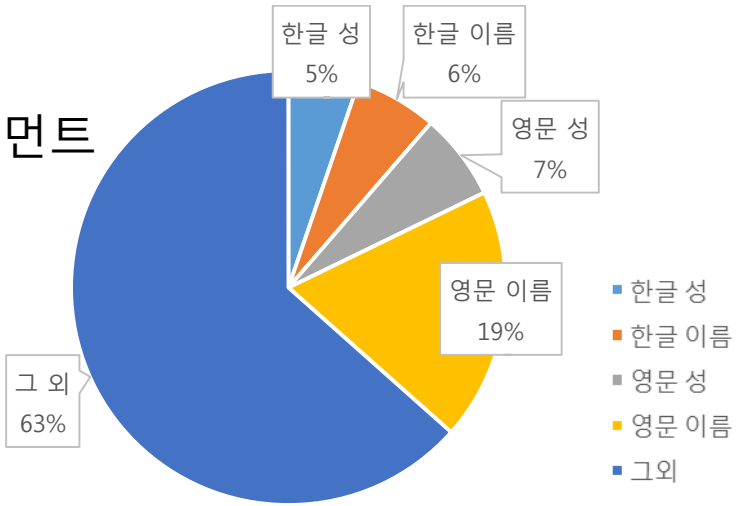
한글 성	한글 이름	영문 성	영문 이름	그 외	끝
5.157%	5.889%	5.624%	17.596%	65.733%	0

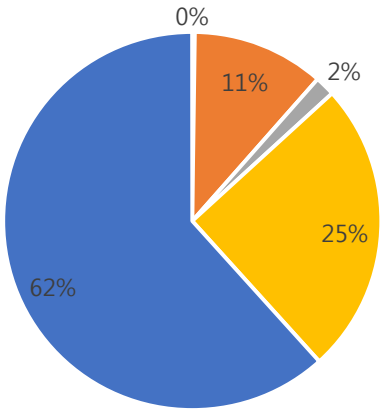
구조적 분석

- HMM (성, 이름만 분리)

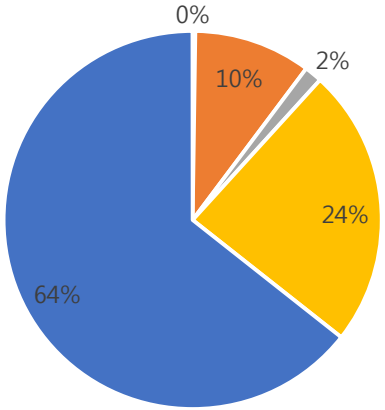
[wjd, whd, rud]	[kor_sur, kor_name, kor_name]
[u, ee]	[eng_name, word]
[da, na]	[eng_name, eng_name]
[dml, duq]	[kor_name, kor_name]
[dl, thdus]	[kor_sur, word]
[my, kym]	[word, word]
[koko]	[word]
[jh]	[word]
[wjd, wp, dbs]	[kor_sur, kor_name, kor_name]
[emfprhs]	[word]
[shfwk]	[word]
[le, no]	[word, word]
[z, ah]	[word, eng_name]
[ckd, ckd]	[kor_name, kor_name]
[lkj]	[word]
[rud, dhks]	[kor_name, kor_name]
[yg]	[word]
[mo, jak, jin]	[eng_name, word, eng_name]
[qkr, tn, dud]	[kor_sur, kor_name, kor_name]
[wns, dlr, cjstk]	[kor_name, kor_name, word]
[bong, j, kim]	[eng_name, word, eng_sur]
[com, kid]	[word, word]
[choi, dong, woo, i]	[eng_sur, eng_name, eng_name, eng_name]
[rotlvkf]	[word]
[mc, hb]	[word, word]
[gus, tjr]	[kor_name, kor_name]
[jpg]	[word]
[cy, d]	[word, word]
[elec, jo, ke]	[word, eng_sur, word]
[seon, bs]	[eng_name, word]

첫 세그먼트





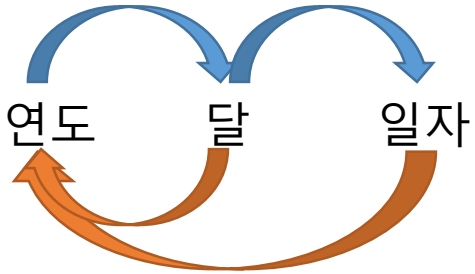
두 번째 세그먼트



세 번째 이후 세그먼트

구조적 분석

- HMM (숫자 분리)



20151113, 151113
112015
11132015

transition matrix 초기값

	연도	달	일자	그 외	끝
연도	0	1	0	0	1
달	1	0	1	0	1
일자	1	0	0	0	1
그 외	0	0	0	1	1
끝	0	0	0	0	1

Start probability 초기값

연도	달	일자	그 외	끝
1	1	0	3	0

- Emission probability
  - ✓ 연도 – 1960, 1961, …, 2015, 60, …, 99, 00, 01, …, 15 에 동일 확률 부여 후 업데이트
  - ✓ 달 – 01, 02, …, 12 에 동일 확률 부여하고 업데이트 없음
  - ✓ 일자 – 01, 02, …, 30 에 동일 확률 부여, 31에 반값 확률 부여, 업데이트 없음
  - ✓ 그 외 state emission probability는 초기값으로 모든 숫자 형태소에 동일한 확률 부여하고, 이후에 업데이트

구조적 분석

- HMM (숫자 분리)
  - 140만개 아이디로 학습

업데이트된 transition probability

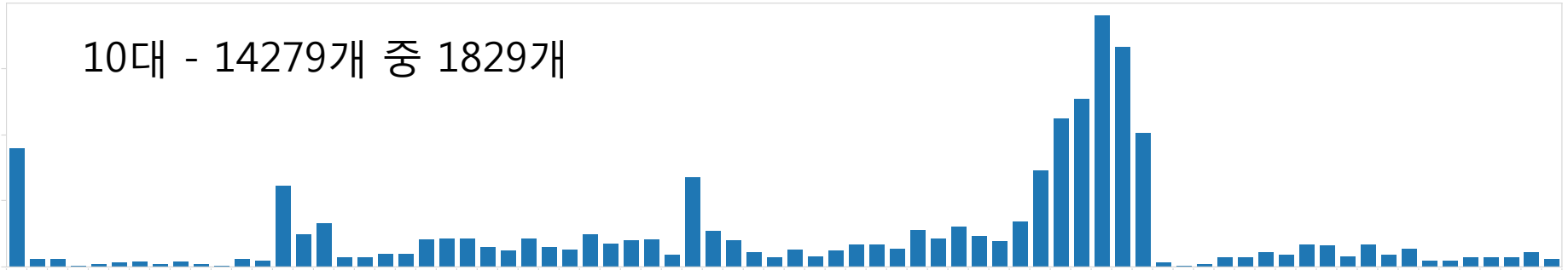
	연도	달	일자	그 외	끝
연도	0	0.214416	0	0	0.785584
달	0.088762	0	0.644677	0	0.266562
일자	0.013316	0	0	0	0.986684
그 외	0	0	0	0.346859	0.653141
끝	0	0	0	0	1

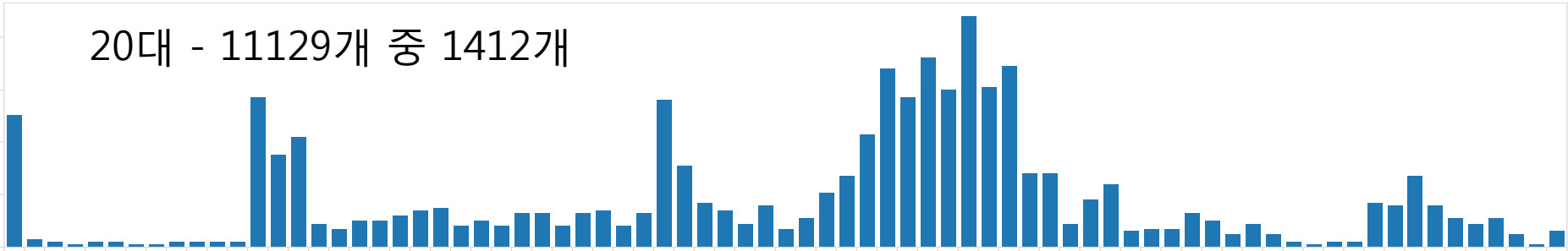
업데이트된 start probability

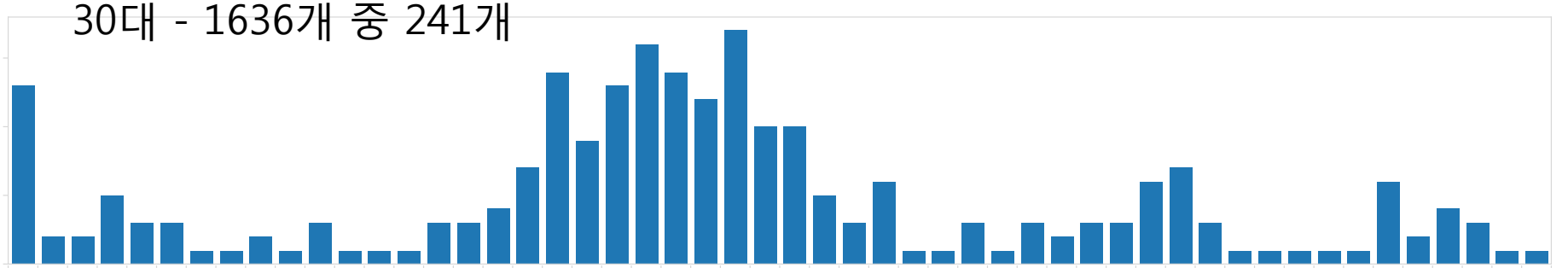
연도	달	일자	그 외	끝
0.180333	0.147644	0	0.672023	0

[96, 08, 10]	[year, month, day]
[14, 567]	[number, number]
[4]	[number]
[754]	[number]
[05, 14]	[month, day]
[971]	[number]
[10, 13]	[month, day]
[96]	[year]
[97, 09, 09]	[year, month, day]
[07]	[month]
[702]	[number]
[12]	[month]
[11, 27]	[month, day]
[010]	[number]
[24, 20]	[number, number]
[157]	[number]
[23, 41]	[number, number]
[82]	[year]
[11, 051]	[number, number]
[72]	[year]
[678]	[number]
[126]	[number]
[2085]	[number]
[35, 24]	[number, number]
[95]	[year]
[99, 87]	[number, number]
[94, 11]	[year, month]
[94]	[year]
[88, 02]	[year, month]
[0192052049]	[number]

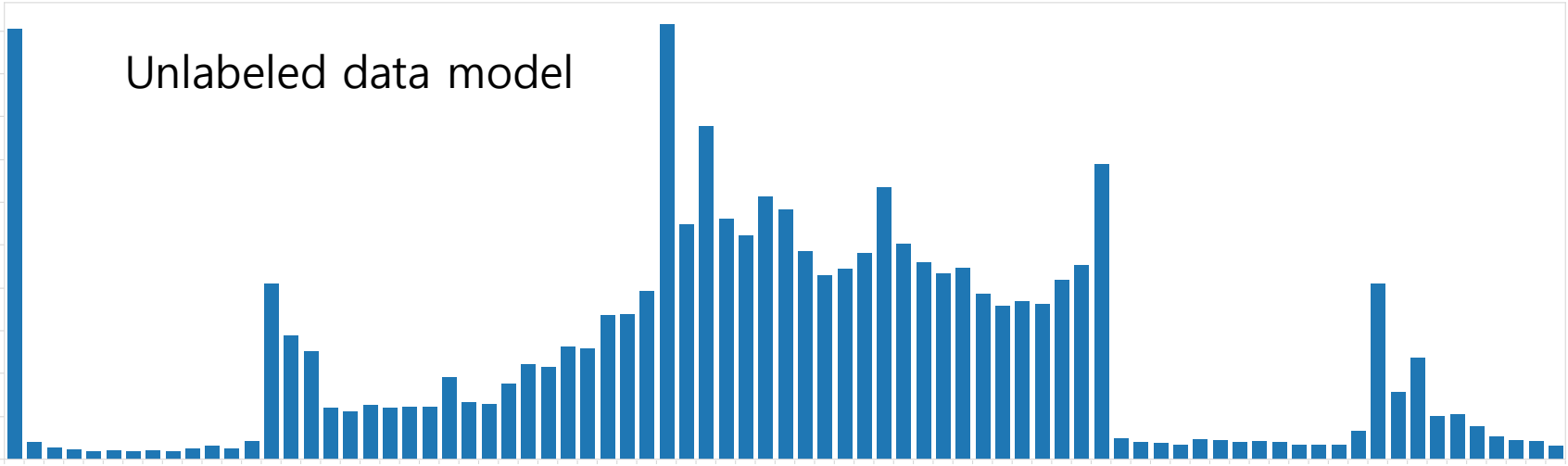
연령대

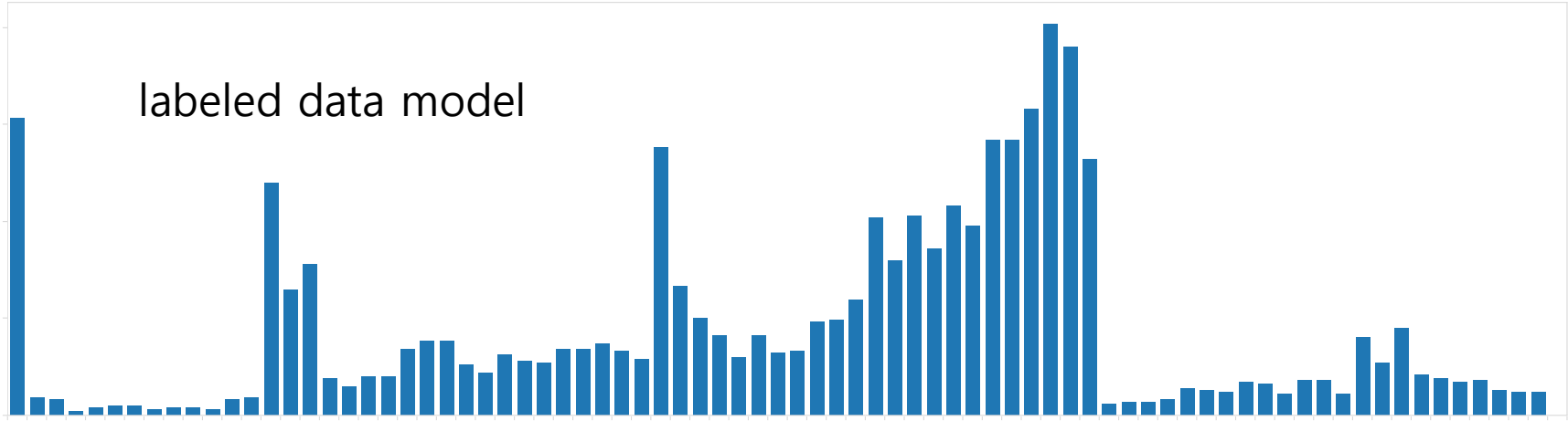






연령대



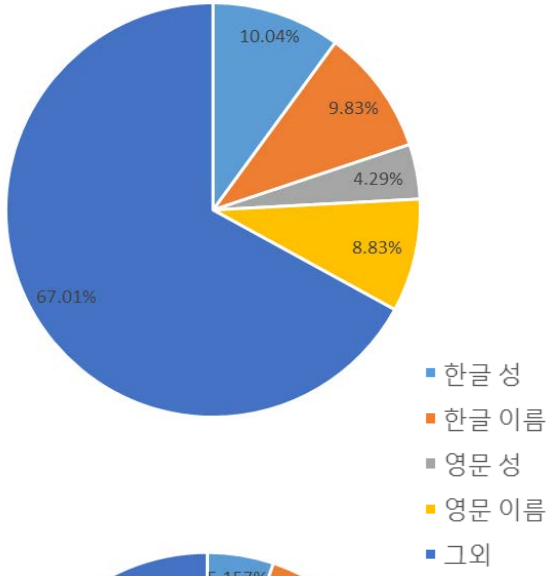


구조적 분석

- HMM 확률 값

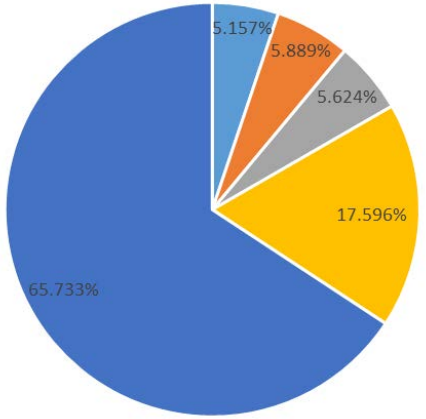
Labeled data 2.7만개

	한글 성	한글 이름	영문 성	영문 이름	그 외	끝
한글 성	0	0.5246	0	0	0.4573	0.0181
한글 이름	0.0050	0.3766	0	0	0.1248	0.4936
영문 성	0	0	0	0.3339	0.4503	0.2158
영문 이름	0	0	0.0092	0.2764	0.2164	0.4980
그 외	0.0030	0.0435	0.0086	0.0437	0.3041	0.5970
끝	0	0	0	0	0	1



Unlabeled data 138만개

	한글 성	한글 이름	영문 성	영문 이름	그 외	끝
한글 성	0	0.5757	0	0	0.4078	0.0165
한글 이름	0.0091	0.3606	0	0	0.1902	0.4401
영문 성	0	0	0	0.2876	0.4206	0.2918
영문 이름	0	0	0.0125	0.2668	0.3258	0.3949
그 외	0.0016	0.0263	0.0161	0.0988	0.4231	0.4341
끝	0	0	0	0	0	1



구조적 분석 기반 분류기

unstructureddataanalysis2015

morfessor

un+structured+data+analysis+2015

hmm

	un	structured	data	analysis	2015
한글 성					
한글 이름					
영문 성					
영문 이름					
그 외					

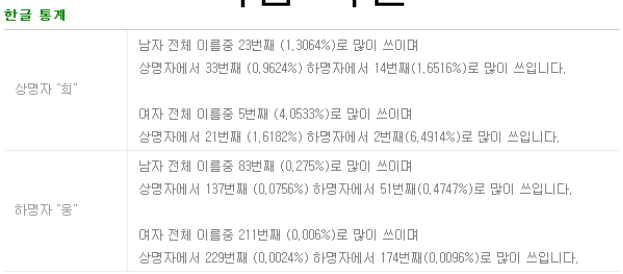
State 별 확률 값

Independent 가정

남/여 likelihood 도출

남/여 분류

이름 확률



한글 통계

상명자 "희"	남자 전체 이름중 23번째 (1,3064%)로 많이 쓰이며 상명자에서 33번째 (0,9624%) 하명자에서 14번째(1,6516%)로 많이 쓰입니다. 여자 전체 이름중 5번째 (4,0533%)로 많이 쓰이며 상명자에서 21번째 (1,6182%) 하명자에서 2번째(6,4914%)로 많이 쓰입니다.
하명자 "웅"	남자 전체 이름중 83번째 (0,275%)로 많이 쓰이며 상명자에서 137번째 (0,0756%) 하명자에서 51번째(0,4747%)로 많이 쓰입니다. 여자 전체 이름중 211번째 (0,006%)로 많이 쓰이며 상명자에서 229번째 (0,0024%) 하명자에서 174번째(0,0096%)로 많이 쓰입니다.

단어 확률

morph	P(m)	P(m|F)	P(m|M)	성향
dms	0.653%	1.073%	0.395%	F
love	0.980%	1.392%	0.726%	F
gml	0.677%	1.042%	0.452%	F
al	0.453%	0.743%	0.274%	F
a	1.685%	2.063%	1.451%	F
hye	0.130%	0.289%	0.032%	F
rhdwn	0.087%	0.217%	0.006%	F
gp	0.197%	0.382%	0.083%	F
wl	0.598%	0.856%	0.439%	F

아이디 구조 분석 테스트

E-mail id	segments	상/하명 분리	이름으로	segments	숫자
hee188	[hee]	[eng_back]	[eng_name]	[188]	[number]
hank	[han, k]	[eng_front, word]	[eng_sur, word]		
hoseong	[ho, seong]	[eng_front, eng_back]	[eng_name, eng_name]		
jinwon	[jin, won]	[eng_front, eng_back]	[eng_name, eng_name]		
misuke88	[mi, suk, e]	[eng_front, eng_back, word]	[eng_name, eng_name, eng_name]	[88]	[year]
wpgur0107	[wp, gur]	[kor_front, kor_back]	[kor_name, kor_name]	[01, 07]	[month, day]
zoon	[zoo, n]	[word, word]	[word, word]		

To do

- To do
  - 구조 분석기 안정화 및 버그 수정
  - 분류기 구축 및 성능 비교
  - 지금까지 발표 자료 정리
  - 논문 작성 준비

- 경청해주셔서 감사합니다.