웹 로그인 아이디와 사용자 성별 및 연령대 관계 분석과 예측 모형
- 형태소 분리와 데이터 탐색

2015. 10. 5.

박희웅

서울대학교 산업공학과

관련 기존 연구

성별 예측

제목	저자	게재지	연도	데이터 소스	사용 피쳐	정확율
Discriminating gender on twitter	Burger et al.	In Proc. EMNLP	2011	twitter.com	screen name	77%
What Your Username Says About You	Jaech and Ostendorf		2015	okcupid.com	username	74%

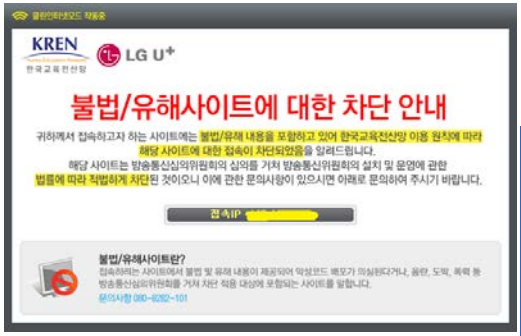
twitter.com

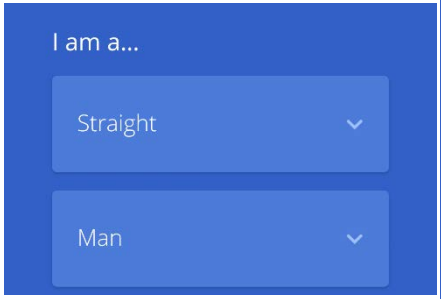
관련 기존 연구

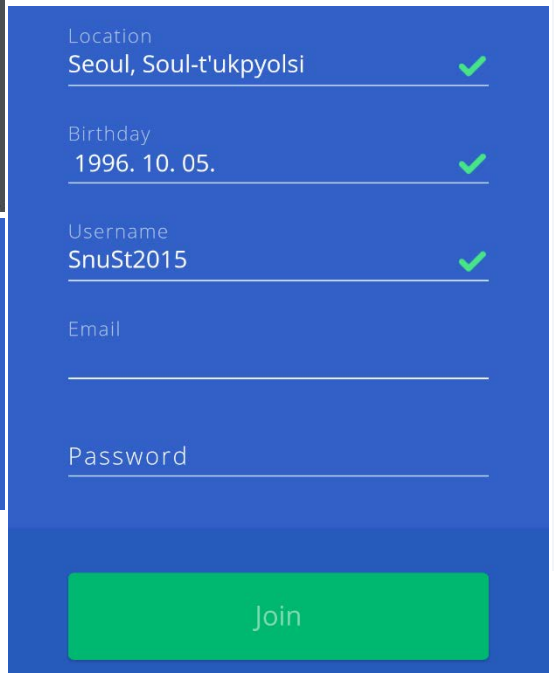
성별 예측

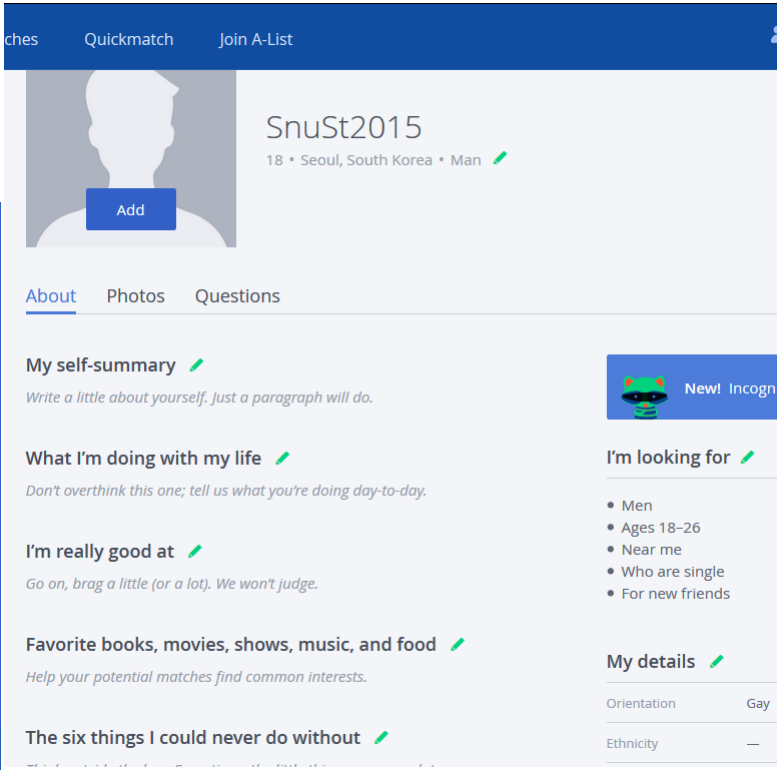
제목	저자	게재지	연도	데이터 소스	사용 피쳐	정확율
Discriminating gender on twitter	Burger et al.	In Proc. EMNLP	2011	twitter.com	screen name	77%
What Your Username Says About You	Jaech and Ostendorf		2015	okcupid.com	username	74%

okcupid.com









Morfessor

- **주요 개념**
  - morpheme: minimal linguistic unit with lexical or grammatical meaning. 우리말로는 형태소
  - u-morphs: 다른 사용자의 아이디와 공유할 수 있을 만큼 작은 단위의 문자열이지만 의미를 보존하고 있음. morpheme과 같은 개념이지만 일반적인 단어에서가 아닌 아이디 문자열으로부터 인코딩된 형태소
- **Morfessor**
  - Creutz and Lagus (2006)
  - Unsupervised morphology induction
  - Maximizing the likelihood of the data and the likelihood of the model라는 두 가지 상반된 목표를 최적화시키는 minimum description length (MDL) 목적함수를 갖고 있음
  - Morfessor software 비영리적으로 사용 가능
    - http://www.cis.hut.fi/projects/morpho
  - Technical paper
    - Creutz, M., & Lagus, K. (2005). Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. In *Helsinki University of Technology*.

Morfessor

- **Maximum a posteriori estimate of the overall probability**
  - The model of language (M) consists of a morph vocabulary, or a *lexicon of morphs*, and a *grammar*
  - Aim at finding the optimal model of language for producing a segmentation of the corpus, i.e., a set of morphs that is concise, and moreover gives a concise representation for the corpus
  - The maximum a posteriori (MAP)

$$\arg\max_{\mathcal{M}} P(\mathcal{M} \mid corpus) = \arg\max_{\mathcal{M}} P(corpus \mid \mathcal{M}) \cdot P(\mathcal{M}), \text{ where} \quad (1)$$

$$P(\mathcal{M}) = P(lexicon, grammar). \quad (2)$$

    - The probability of the model of language P(M)
    - The maximum likelihood (ML) estimate of the corpus conditioned on the given model of language, written as P(corpus |M)

Morfessor

- **Lexicon**
  - The lexicon contains one entry for each distinct morph (morph type) in the segmented corpus
  - an inventory of whatever information one might want to store regarding a set of morphs, including their interrelations
  - the probability of coming up with a particular set of M morphs making up the lexicon

$$P(lexicon) = M! \cdot P(properties(\mu_1), \ldots, properties(\mu_M)). \quad (3)$$

  - In the Baseline versions of Morfessor, two properties are stored
    - the frequency (number of occurrences) of the morph in the corpus
    - the string of letters that the morph consists of. This property contains knowledge about the length of the morph, i.e., the number of letters in the string
  - assume that the frequency and morph string values are independent of each other

$$P(properties(\mu_1), \ldots, properties(\mu_M)) = P(f_{\mu_1}, \ldots, f_{\mu_M}) \cdot P(s_{\mu_1}, \ldots, s_{\mu_M}), \quad (4)$$

Morfessor

- **Grammar**
  - information about how language units can be combined, i.e. which morphs precede or follow it or whether the morph is placed in the beginning, middle, or end of a word
  - In the Baseline models, no context-sensitivity. The probability P(lexicon, grammar) reduces to P(lexicon)

- **Corpus**
  - sequence of some morphs that are present in the lexicon
  - In MAP modeling, the one most probable segmentation is chosen
  - the probability of the corpus, when a particular model of language (lexicon and non-existent grammar)

$$P(\mu_i) = \frac{f_{\mu_i}}{N} = \frac{f_{\mu_i}}{\sum_{j=1}^{M} f_{\mu_j}}. \tag{5}$$

$$P(corpus \mid \mathcal{M}) = \prod_{j=1}^{W} \prod_{k=1}^{n_j} P(\mu_{jk}). \tag{6}$$

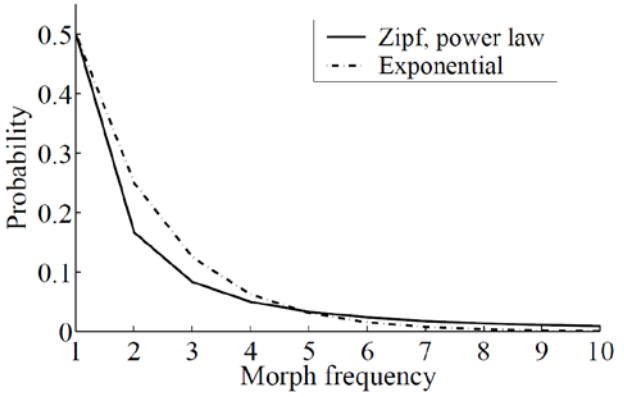
Morfessor

- **Properties of the morphs in the lexicon: Frequency**
- **Implicit model**
  - *non-informative prior*

$$P(f_{\mu_1}, \ldots, f_{\mu_M}) = 1 / \binom{N-1}{M-1} = \frac{(M-1)!(N-M)!}{(N-1)!}, \qquad (7)$$

  - where N is the total number of morph tokens in the corpus, which equals the sum of the frequencies of the M morph types that make up the lexicon
- **Explicit model**
  - assigns a particular probability to every possible morph frequency value
  - assume that the frequency of one morph is independent of the frequencies of the other morphs
  - an expression based on Zipf's law

$$P(f_{\mu_i}) = f_{\mu_i}^{\log_2(1-h)} - (f_{\mu_i} + 1)^{\log_2(1-h)}.$$



Morfessor

- **Properties of the morphs in the lexicon: Length**
- **Implicit model**
  - a morph consists of is independent of the strings that the other morphs consist of

$$P(s_{\mu_1}, \ldots, s_{\mu_M}) = \prod_{i=1}^{M} P(s_{\mu_i}). \quad (10)$$

  - assume that the letters in a morph string are drawn from a probability distribution independently of each other

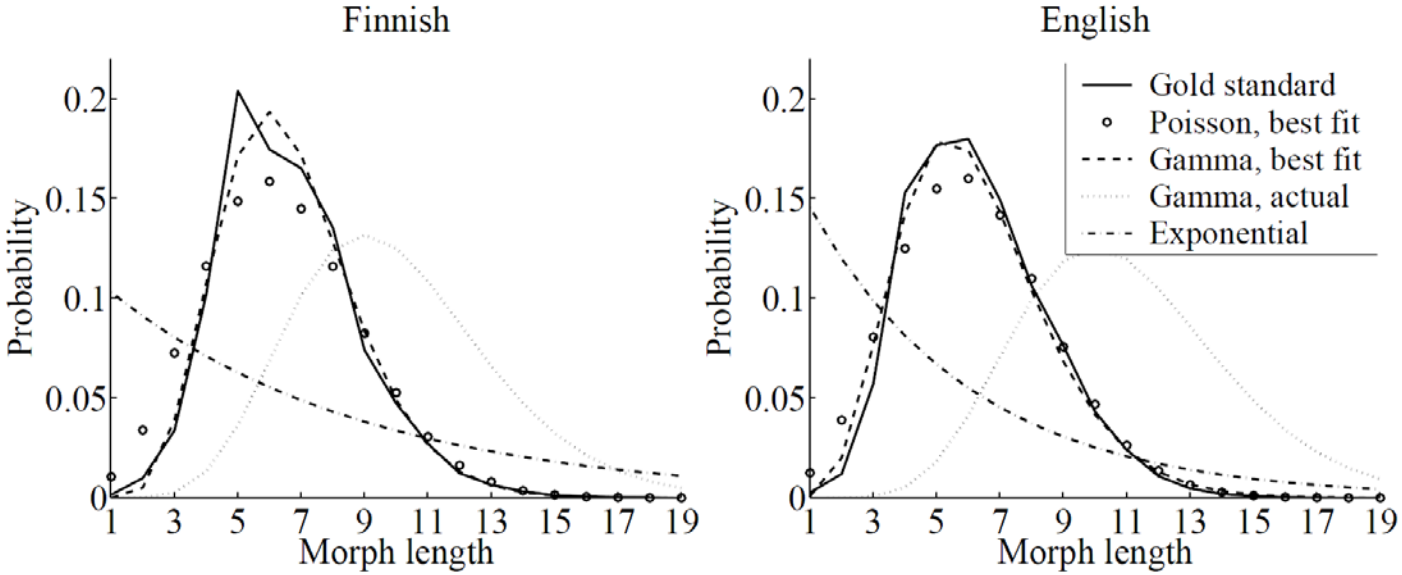
$$P(s_{\mu_i}) = \prod_{j=1}^{l_{\mu_i}} P(c_{ij}). \quad (11)$$

  - The probability distribution over the alphabet $P(c_{ij})$ is estimated from the corpus by computing relative frequencies of each of the letters observed
  - a special *end-ofmorph* character that is part of the alphabet and is appended to each morph string in the lexicon and marks the end of the string. It implies an *exponential distribution*

$$P(l) = [1 - P(\#)]^l \cdot P(\#), \quad (12)$$

Morfessor

- **Properties of the morphs in the lexicon: Length**
- **Explicit model**
  - Instead of using an end-of-morph marker for the morphs in the lexicon, one can first decide the length of the morph according to an appropriate probability distribution and then choose the selected number of letters
  - model the length distribution of *morph types* (the morphs in the lexicon)
  - use a *gamma distribution*



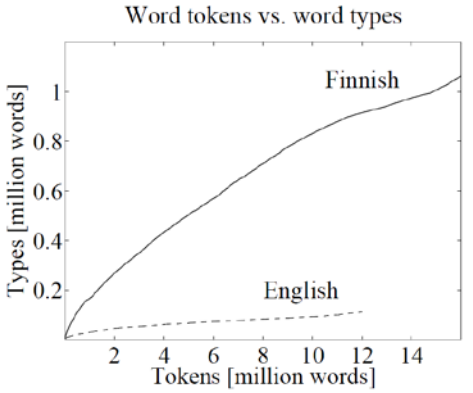
Morfessor

- **Experiments**

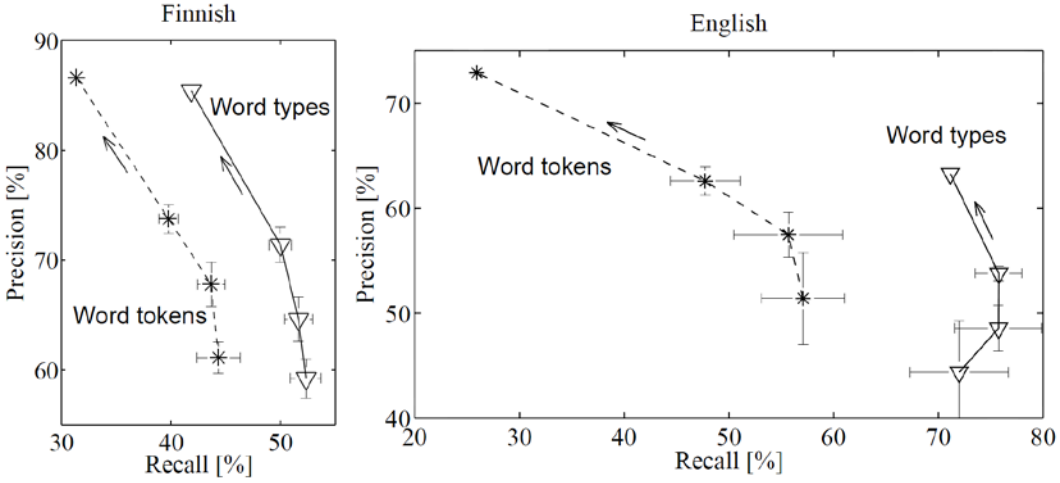
Agglutinative language (교착어):
Algonquian, namely Cree and Blackfoot
Japanese, Korean, Mongolic,
Tungusic, Turkic
Armenian, Athabaskan, Austronesian
Bantu, Eskimo–Aleut , namely Aleut,
Inuktitut, and Yupik
many Uralic , namely Hungarian, Finnish
and Sami, etc.

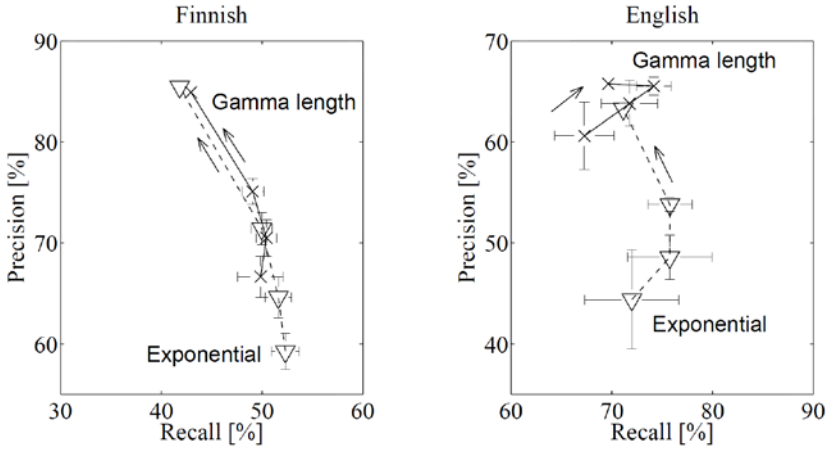
Corpus of Finish and English

word tokens	Finnish word types	English word types
10 000	5 500	2 400
50 000	20 000	7 200
250 000	65 000	17 000
12 000 000	–	110 000
16 000 000	1 100 000	–



Learning a morph lexicon from
word tokens vs. word types

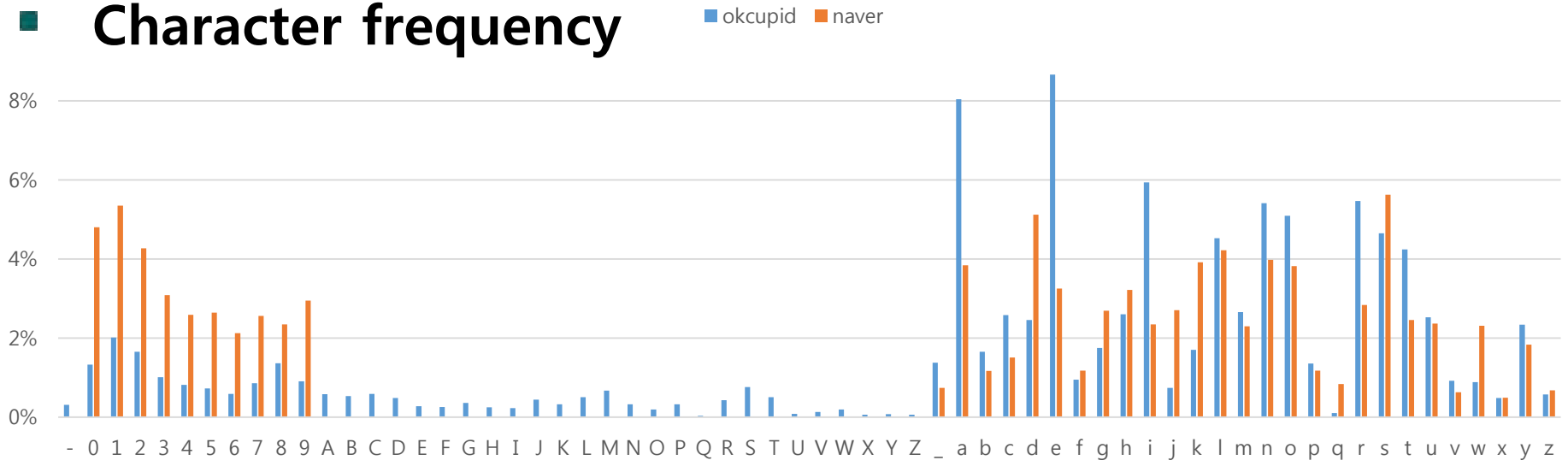


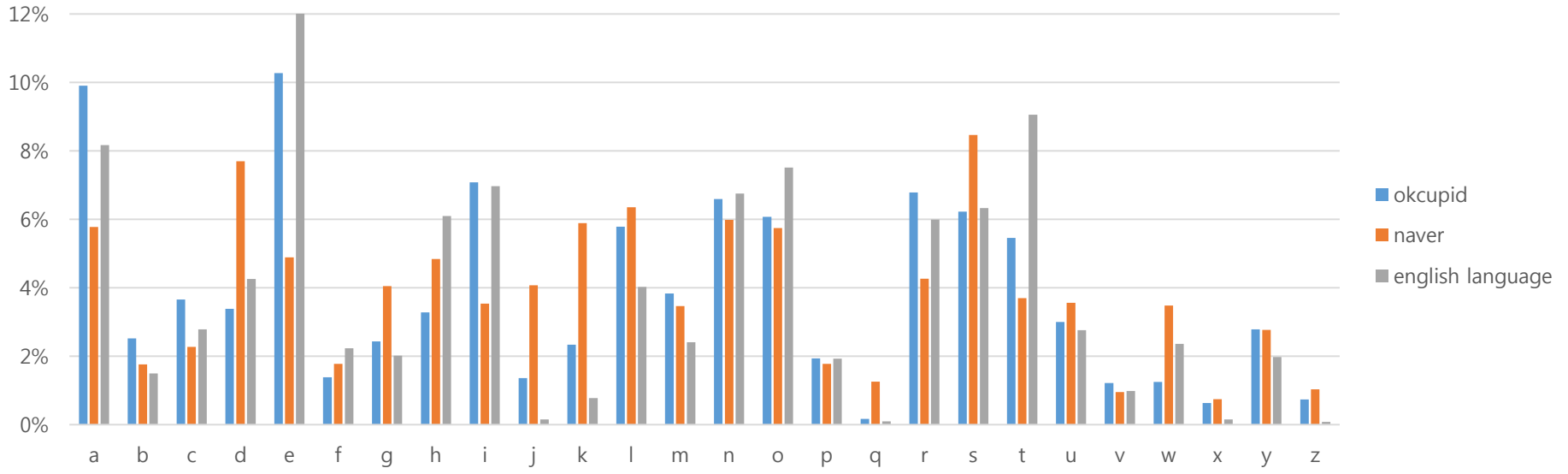


Evaluation of
the explicit gamma length prior

Data visualization

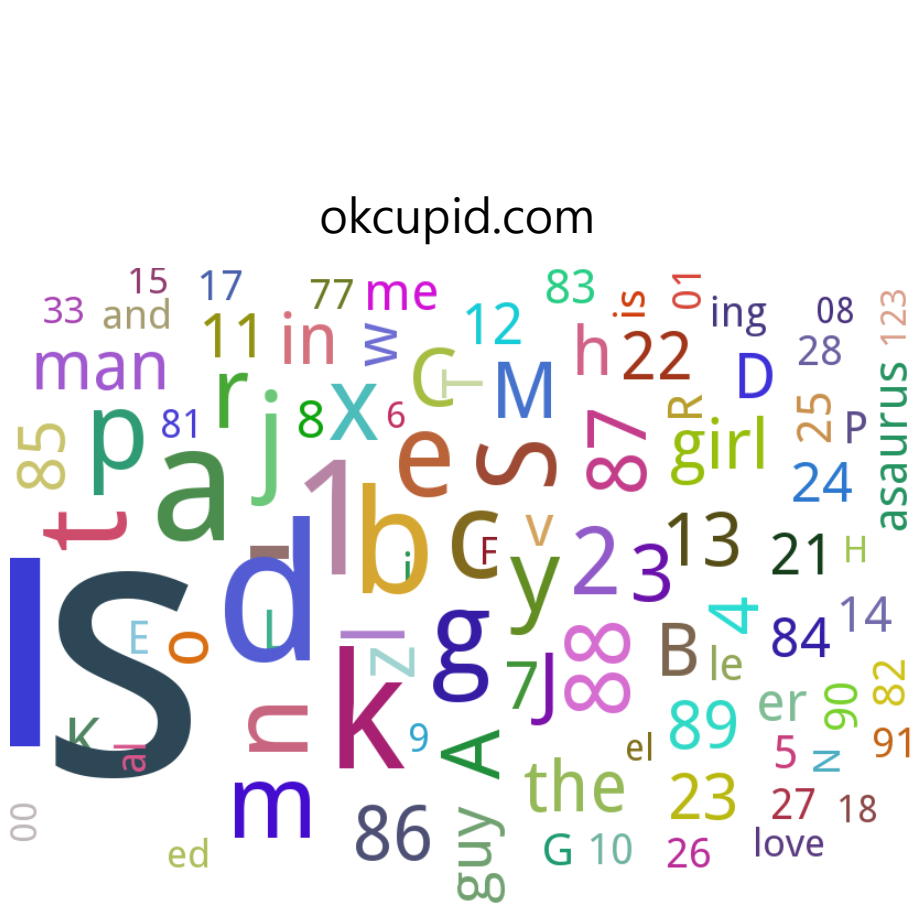
- **Character frequency**

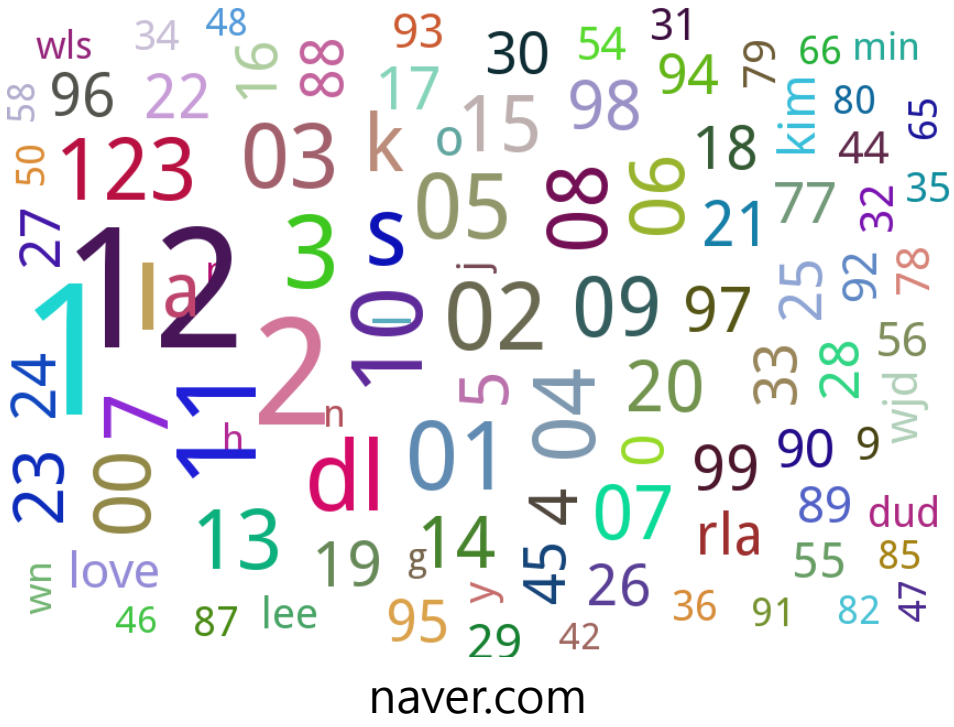




Data visualization

- 빈도수 상위 morpheme





Data visualization

- 빈도수 상위 morpheme (알파벳으로 구성된)

okcupid.com

naver.com

EOD

- **To do**
  - 남녀 성별과 연령대로 구분하여 판별력 높은 morph 탐색
  - Baseline classifier (적군) 분석
  - 한영 변환, 연결 알고리즘 조사
  - 중간 보고서 준비

- **참고 문헌**

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating Gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1301–1309). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2145432.2145568

Jaech, A., & Ostendorf, M. (2015). What Your Username Says About You. *rXiv:1507.02045 [cs]*. Retrieved from http://arxiv.org/abs/1507.02045.

Creutz, M., & Lagus, K. (2005). Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. In *Helsinki University of Technology*.

Creutz, M., & Lagus, K. (2006). Morfessor in the morpho challenge. In *Proceedings of the PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*.

Virpioja, S., Smit, P., Grönroos, S.-A., & Kurimo, M. (2013). *Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline*. Aalto University. Retrieved from https://aaltodoc.aalto.fi:443/handle/123456789/11836