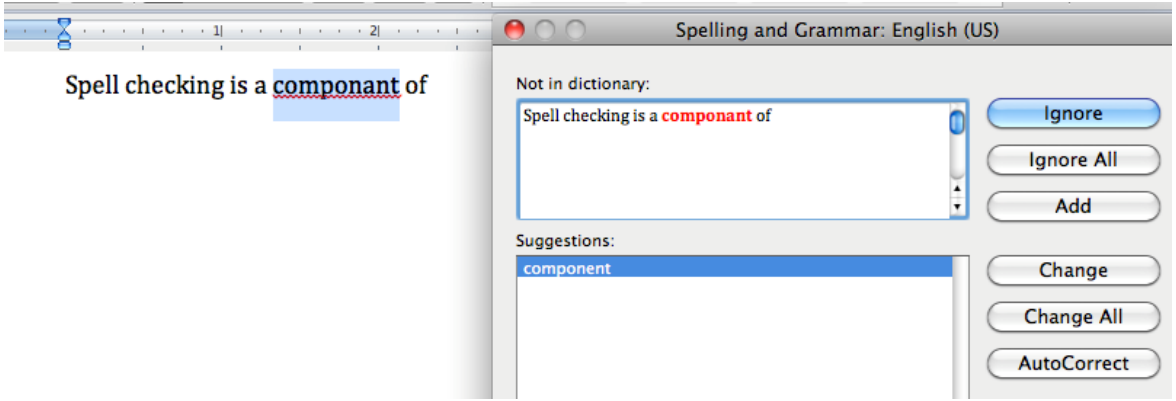
Үгийн алдаа зүгшрүүлэлт ба шуугиант суваг

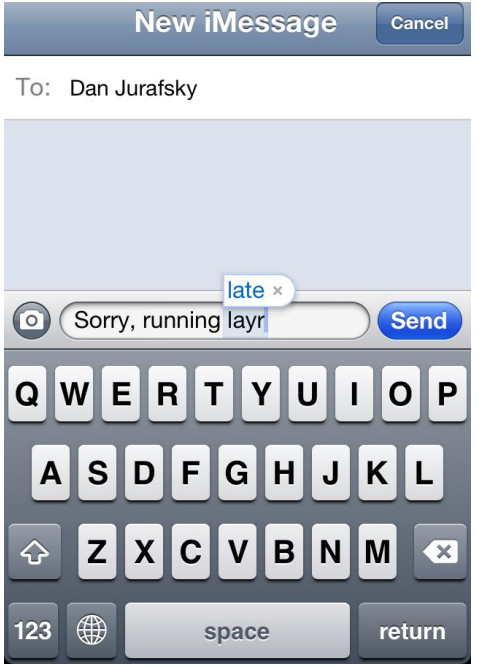
Үгийн алдаа зүгшрүүлэх бодлого

Үгийн алдаа зүгшрүүлдэг программууд

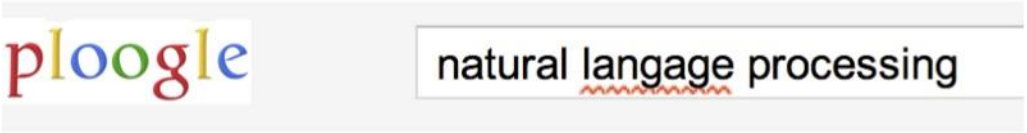
Word програм



Гар утас



Web search



Showing results for natural ***language*** processing

Search instead for natural langage processing

Зөв бичгийн дүрмийн даалгавар

- Зөв бичгийн алдаа таних
- Зөв бичгийн алдааг зүгшрүүлэх:
  - Автоматаар засах
    - hte→the
  - Зөв нэгийг санал болгох
  - Жагсаалтаар санал болгох

Зөв бичгийн алдааны төрлүүд

- Буруу бичсэн үгийн алдаа - Non-word Errors
  - *graffe →giraffe*
- Зөв бичиглэлтэй үгийн алдаа - Real-word Errors
  - Хэвлэлийн алдаа - Typographical errors
    - *three →there*
  - Танин мэдэхүйн алдаа - Cognitive Errors (ижил дуудлага - homophones)
    - *piece→peace,*
    - *too → two*

Зөв бичгийн алдааны үзүүлэлт

26%: Веб хайлтын үр дүн Wang *et al.* 2003

13%: backspace товч ашиглахгүй бол 13% алдаа гаргадаг: Whitelaw *et al.* English&German

7%: жижиг төхөөрөмж дээр 7% үгийн алдаа засдаг

2%: гэхдээ 2% алдаа засагдаагүй үлддэг Soukoreff &MacKenzie 2003

1-2%: Компьютерийн гараар ийм хувийг дахиж бичдэг: Kane and Wobbrock 2007, Gruden et al. 1983

Бүруу бичсэн үгийн алдаа

- Буруу бичсэн үгийн алдааг таних:
  - ***Толь бичигт*** байхгүй ямар ч үг алдаа
  - Том толь бичиг байвал сайн
- Буруу бичсэн үгийн алдааг зүгшрүүлэх:
  - ***Санал болгох үгсийг*** үүсгэх: алдаатай үгтэй төстэй бодит үгс
  - Хамгийн тохирохыг сонгох:
    - Хамгийн богино, хамгийн бага, жигнэсэн засварын хэмжээ
    - Хамгийн өндөр шуугиант сувгийн магадлал

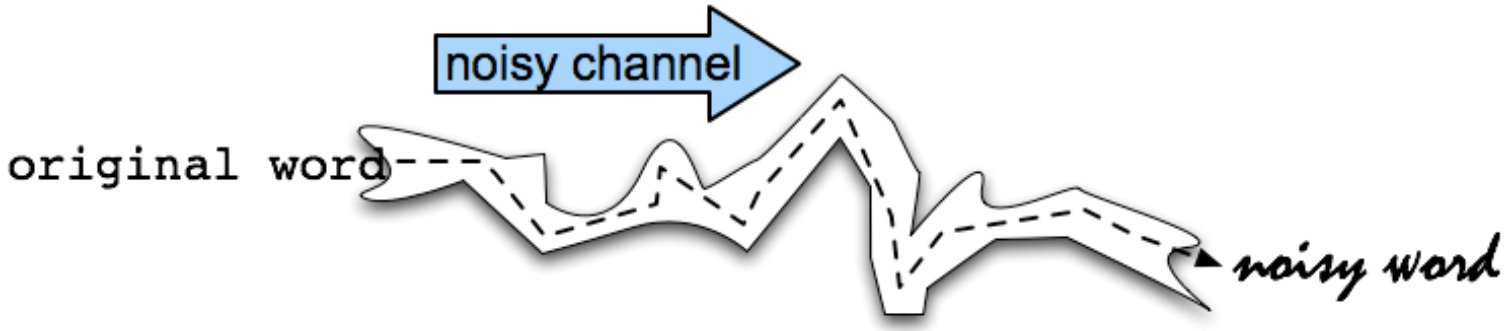
Зөв бичиглэлтэй үгийн алдаа

- *w* үг бүрийн хувьд, санал болгох үгийн олонлогийг үүсгэх:
  - санал болгох ***ижил дуудлагатай үгсийг*** хайх
  - санал болгох ***ижил бичлэгтэй үгсийг*** хайх
  - Санал болгох жагсаалтад *w* үгийг өөрийг нь оруулах
- Хамгийн сайн тохирох үгийг сонгох
  - Шуугиант суваг
  - Ангилагч

Үгийн алдаа зүгшрүүлэлт ба шуугиант суваг

Зөв бичгийн шуугиант сувгийн загвар

Шуугиант сувгийн санаа



Шуугиант суваг – магадлалын загвар

- Алдаатай бичсэн x үгийн ажиглалтыг харж байна
- зөв үг w –г хай

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} P(w \mid x)$$

$$= \underset{w \in V}{\operatorname{argmax}} \frac{P(x \mid w)P(w)}{P(x)}$$

Bayes-ийн дүрэм

$$= \underset{w \in V}{\operatorname{argmax}} P(x \mid w)P(w)$$

Түүх: 1990 оны үед зөв бичгийн алдааг засахад санал болгосон шуугиант суваг

- **IBM**
  - Mays, Eric, Fred J. Damerau and Robert L. Mercer. 1991. Context based spelling correction. *Information Processing and Management*, 23(5), 517–522
- **AT&T Bell Labs**
  - Kernighan, Mark D., Kenneth W. Church, and William A. Gale. 1990. A spelling correction program based on a noisy channel model. Proceedings of COLING 1990, 205-210

Буруу бичсэн үгийн алдааны жишээ

acress

Санал болгох үгсийн үүсгэлт

- Төстэй бичигддэг үгс
  - бага засварын хэмжээ
- Ижил дуудлагатай үгс
  - дуудлагын бага засварын хэмжээ

Damerau-Levenshtein засварын хэмжээ

- Засварлаж байгаа 2 тэмдэгт мөр хоорондын хамгийн бага засварын хэмжээ:
  - Оруулалт
  - Арилгалт
  - Орлуулалт
  - Зэргэлдээ 2 үсгийн байр солилт

`acress` үгтэй 1 засварын хэмжээтэй үгс

Алдаа	Санал болгох зүгшрүүлэлт	Зүгшрүүлэх үсэг	Алдаатай үсэг	Төрөл
acress	actress	t	–	хасалт
acress	cress	–	a	оруулалт
acress	caress	ca	ac	байр солилт
acress	access	c	r	орлуулалт
acress	across	o	e	орлуулалт
acress	acres	–	s	оруулалт
acress	acres	–	s	оруулалт

Санал болгох үгсийн үүсгэлт

- алдааны 80% нь засварын хэмжээ 1 дотор байдаг
- Бараг бүх алдаа засварын хэмжээ 2 –оос хэтэрдэггүй
- Мөн **зай** эсвэл **дундуур зураас** оруулахыг зөвшөөрдөг
  - `thisidea` → `this idea`
  - `inlaw` → `in-law`

Хэлний загвар

- Мэддэг бүх хэлний загварчлалын дурын алгоритмыг ашигла
- Юниграм, биграм, триграм
- Веб хэмжээт зөв бичгийн алдаа зүгшрүүлэлт
  - Ухаангүй буцаж шилжих

Өмнөх юниграм магадлал

Corpus of Contemporary English (COCA) дахь 404,253,213 үгсээс тоолсон

үг	үгийн давтамж	P(үг) - магадлал
actress	9,321	.0000230573
cress	220	.0000005442
caress	686	.0000016969
access	37,038	.0000916207
across	120,844	.0002989314
acres	12,874	.0000318463

Сувгийн загварын магадлал

- **Алдааны загварын магадлал, засварлах магадлал**
- *Kernighan, Church, Gale 1990*

- *алдаатай бичсэн үг* $x = x_1, x_2, x_3 \ldots x_m$
- *зөв үг* $w = w_1, w_2, w_3, \ldots, w_n$

- P(x|w) = засварын магадлал
  - (арилгалт/оруулалт/орлуулалт/байр солилт)

Алдааны магадлалыг тооцоолох: андуурлын матрици

```
del[x,y]:    count(xy typed as x)
ins[x,y]:    count(x typed as xy)
sub[x,y]:    count(x typed as y)
trans[x,y]:  count(xy typed as yx)
```

өмнөх тэмдэгт дээр суурилсан оруулалт болон арилгалт

Зөв бичгийн алдааны андуурлын матрици

sub[X, Y] = Substitution of X (incorrect) for Y (correct)

X \ Y (correct)	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	0
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	15	0	1	0	18	0
f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	0
p	0	11	1	2	0	6	5	0	2	9	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	0
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	0
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0

Андуурлын матрици үүсгэх нь

- Peter Norvig’s list of errors
- Peter Norvig’s list of counts of single-edit errors

Сувгийн загвар

Kernighan, Church, Gale 1990

$$P(x|w) = \begin{cases} \frac{\text{del}[w_{i-1}, w_i]}{\text{count}[w_{i-1} w_i]}, & \text{if deletion} \\ \frac{\text{ins}[w_{i-1}, x_i]}{\text{count}[w_{i-1}]}, & \text{if insertion} \\ \frac{\text{sub}[x_i, w_i]}{\text{count}[w_i]}, & \text{if substitution} \\ \frac{\text{trans}[w_i, w_{i+1}]}{\text{count}[w_i w_{i+1}]}, & \text{if transposition} \end{cases}$$

acress –н хувьд сувгийн загвар

Санал болгох зүгшрүүлэлт	зүгшрүүлэх үсэг	алдаатай үсэг	x|w	P(x|word)
actress	t	-	c|ct	.000117
cress	-	a	a|#	.00000144
caress	ca	ac	ac|ca	.00000164
access	c	r	r|c	.000000209
across	o	e	e|o	.0000093
acres	-	s	es|e	.0000321
acres	-	s	ss|s	.0000342

acress –н хувьд шуугиант сувгийн магадлал

Санал болгох зүгшрүүлэлт	зүгшрүүлэх үсэг	алдаатай үсэг	x|w	P(x|word)	P(word)	10^9 *P(x|w)P(w)
actress	t	-	c|ct	.000117	.0000231	2.7
cress	-	a	a|#	.00000144	.000000544	.00078
caress	ca	ac	ac|ca	.00000164	.00000170	.0028
access	c	r	r|c	.000000209	.0000916	.019
across	o	e	e|o	.0000093	.000299	2.8
acres	-	s	es|e	.0000321	.0000318	1.0
acres	-	s	ss|s	.0000342	.0000318	1.0

acress –н хувьд шуугиант сувгийн магадлал

Санал болгох зүгшрүүлэлт	зүгшрүүлэх үсэг	алдаатай үсэг	x|w	P(x|word)	P(word)	10^9 *P(x|w)P(w)
actress	t	-	c|ct	.000117	.0000231	2.7
cress	-	a	a|#	.00000144	.000000544	.00078
caress	ca	ac	ac|ca	.00000164	.00000170	.0028
access	c	r	r|c	.000000209	.0000916	.019
across	**o**	**e**	**e|o**	**.0000093**	**.000299**	**2.8**
acres	-	s	es|e	.0000321	.0000318	1.0
acres	-	s	ss|s	.0000342	.0000318	1.0

Үнэлгээ

- Зарим зөв бичгийн алдаа шалгах олонлог
  - Wikipedia’s list of common English misspelling
  - Aspell filtered version of that list
  - Birkbeck spelling error corpus
  - Peter Norvig’s list of errors (includes Wikipedia and Birkbeck, for training or testing)