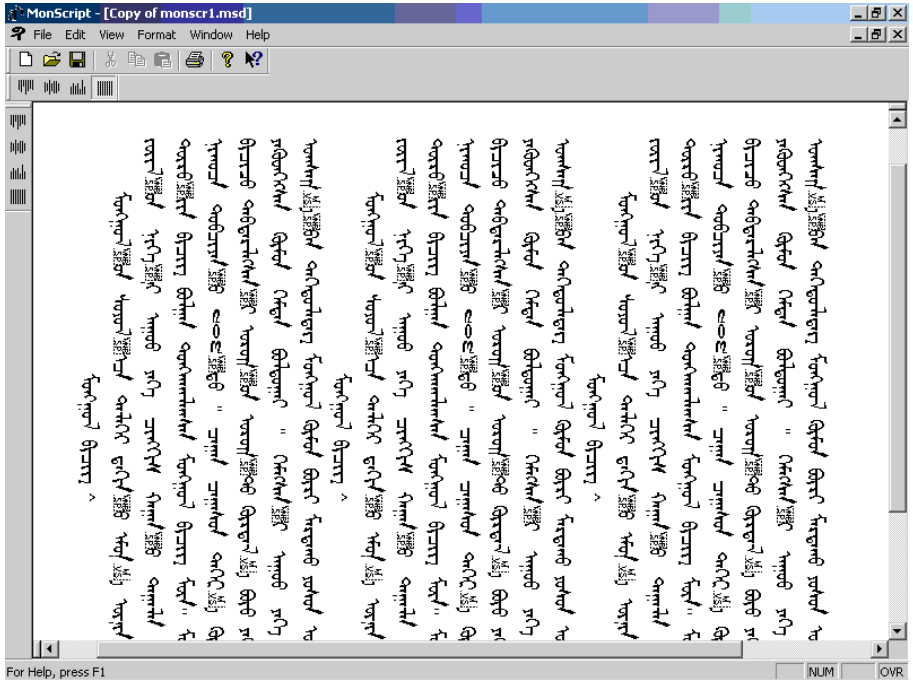


Текст ангилал ба Найв Бэйс

Текст ангилал: Үнэлгээ

2-оос олон ангийн хувьд: 2-тын ангилагчийн олонлогууд

- нэг эсвэл олон утгат ангилагч
  - Нэг документ 0, 1, эсвэл 1-ээс олон ангид хамаарч болно.

- For each анги $c \in C$
  - c-г бусад $c' \in C$ бүх ангиудаас ялгахын тулд нэг γ_c ангилагчийг үүсгэнэ.
- Өгөгдсөн тест документ d –ын хувьд,
  - γ_cбүрийг ашиглаж байгаа анги бүр доторх гишүүн бүрийн хувьд түүнийг үнэлнэ.
  - d нь true буцаадаг γ_c –ийн хувьд ямар ч ангид хамаарна.

2-оос олон ангийн хувьд: 2-тын ангилагчийн олонлогууд

- Яг нэг эсвэл олон гишүүнт ангилал
  - Ангиуд нь бие биеэсээ ялгаатай: документ бүр яг нэг ангид байна.

- For each анги $c \in C$
  - c-г бусад $c' \in C$ бүх ангиудаас ялгахын тулд нэг γ_c ангилагчийг үүсгэнэ.
- Өгөгдсөн тест документ d –ын хувьд,
  - γ_cбүрийг ашиглаж байгаа анги бүр доторх гишүүн бүрийн хувьд түүнийг үнэлнэ.
  - d нь хамгийн өндөр үнэлгээтэй нэг ангид хамаарна.

Үнэлгээ:
Reuters-21578 сонгодог өгөгдлийн олонлог

- Хамгийн их ашигладаг өгөгдлийн олонлог, 21,578 док (90 төрөл бүр, 200 токентой)
- 9603 сургалтын, 3299 тестийн нийтлэл
- 118 категори
  - Нэг нийтлэл 1-ээс олон категорид байж болно.
  - Сургалтын 118 хоёртын ялгаатай категори
- Документ дунджаар (багадаа 1 ангид) 1.24 ангид байна.
- 118 категориос ойролцоогоор 10 нь л том категори

Ерөнхий категориуд
(#train, #test)

- Earn (2877, 1087)
- Acquisitions (1650, 179)
- Money-fx (538, 179)
- Grain (433, 149)
- Crude (389, 189)
- Trade (369,119)
- Interest (347, 131)
- Ship (197, 89)
- Wheat (212, 71)
- Corn (182, 56)

Reuters-н текст категорчлолын өгөгдлийн олонлогийн (Reuters-21578) документ

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981" NEWID="798">

<DATE> 2-MAR-1987 16:51:43.42</DATE>

<TOPICS><D>livestock</D><D>hog</D></TOPICS>

<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>

<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Pork Congress kicks off tomorrow, March 3, in Indianapolis with 160 of the nations pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC.

Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter

&#3;</BODY></TEXT></REUTERS>

с андуурлын матрици

- <c_1,c_2> хос анги бүрийн хувьд c_2луу буруу хуваарилагдсан c_1 –ийн хэдэн документ байна вэ ?
  - $c_{3,2}$: poultry руу бүруу хуваарилагдсан 90 wheat –н документ

тестийн олонлог дах докууд	UK руу хуваарил	poultry руу	wheat руу хуваарил	coffee руу хуваарил	interest хуваарил	trade руу хуваарил
True UK	95	1	13	0	1	0
True poultry	0	1	0	0	0	0
True wheat	10	90	0	1	0	0
True coffee	0	0	0	34	3	7
True interest	-	1	2	13	26	5
True trade	0	0	2	14	5	10

Анги бүрээр үнэлгээний хэмжээс

Recall:

i анги дах зөв ангилсан докуудын хувь:

$$\frac{c_{ii}}{\sum_j c_{ij}}$$

Precision:

i ангид байсан докууд *i* анги гэж хуваарилагдсан докуудын хувь:

$$\frac{c_{ii}}{\sum_j c_{ji}}$$

Accuracy: (1 – алдааны үзүүлэлт)

Зөв ангилсан докуудын хувь:

$$\frac{\sum_i c_{ii}}{\sum_j \sum_i c_{ij}}$$

Микро ба Макро дундажлалт

- Хэрэв бидэнд нэгээс олон анги байвал, олон гүйцэтгэлийн хэмжээсийг хэрхэн нэг тоон хэмжээ рүү нийлүүлэх вэ?
- **Макро дундажлалт**: Анги бүрийн гүйцэтгэлийг тооцоолж, дараа нь дундажлах.
- **Микро дундажлалт**: бүх ангийн хувьд үр дүнг цуглуулж, санамсаргүйн хүснэгтийг тооцоолж үнэлэх.

Микро ба Макро дундажлалт:Жишээ

Анги 1

	үнэн: yes	үнэн: no
Ангилагч: yes	10	10
Ангилагч: no	10	970

Анги 2

	үнэн: yes	үнэн: no
Ангилагч: yes	90	10
Ангилагч: no	10	890

Микро.дунд.хүснэгт

	үнэн: yes	үнэн: no
Ангилагч: yes	100	20
Ангилагч: no	20	1860

- Макро дундажласан precision: (0.5 + 0.9)/2 = 0.7
- Микро дундажласан precision: 100/120 = .83
- Микро дундажласан оноо нь нийтлэг ангиуд дээрх үнэлгээгээр хуваасантай тэнцүү

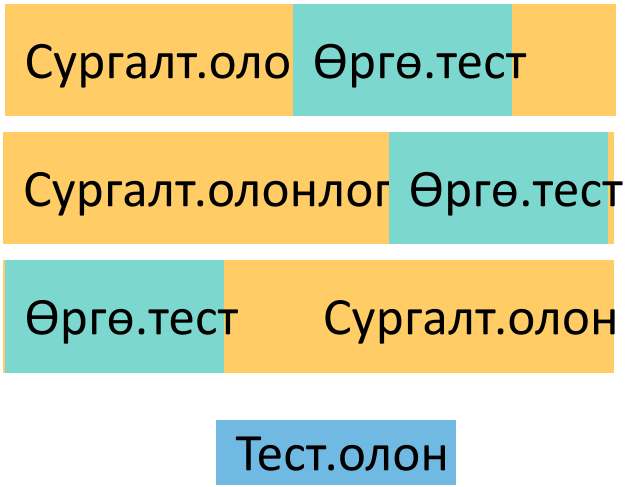
Өргөтгөсөн тестийн олонлогууд ба хөндлөнгийн үнэлгээ

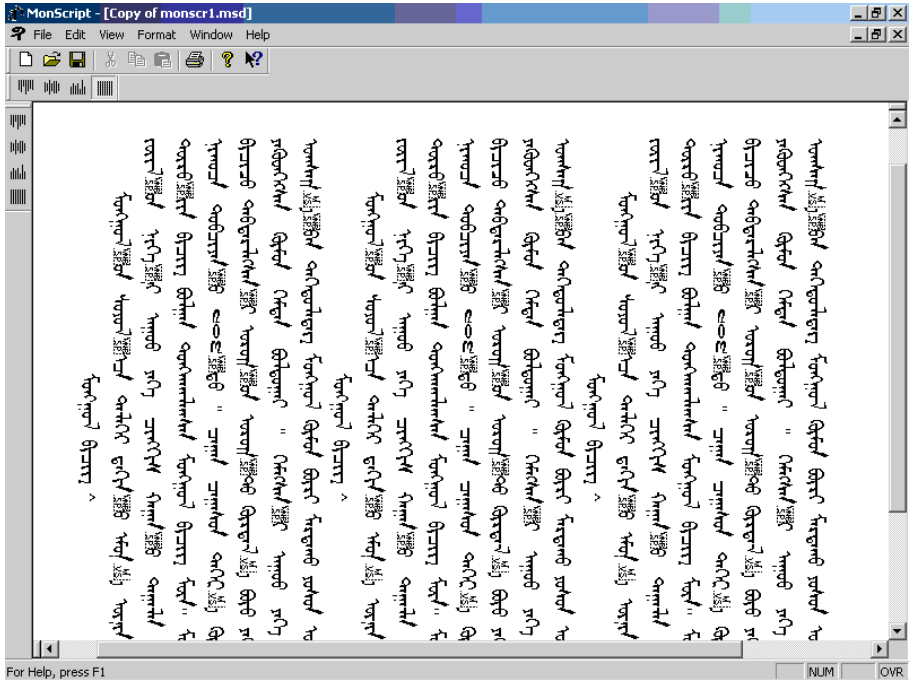
Сургалтын олонлог

Өргөтгөсөн тестийн олонлог

Тестийн олонлог

- Хэмжигдэхүүн: P/R/F1 эсвэл Accuracy
- Үзэгдээгүй тестийн олонлог
  - хэт тохируулахаас зайлсхий (‘тестийн олонлогийг тааруулах’)
  - гүйцэтгэлийн илүү хуучны үнэлгээ
- Олон хуваалт дээрх хөндлөнгийн үнэлгээ
    - ялгаатай өгөгдлийн олонлогуудаас түүврийн алдаануудыг зохицуулах
  - Хуваалт бүр дээрх үр дүнгүүдийг нэгтгэнэ
  - Өргөтгөсөн олонлогын нэгтгэсэн гүйцэтгэлийг тооцоолох





Текст ангилал ба Найв Бэйс

Текст ангилал: Практик асуудлууд

Бодит байдалд

- Ий, Би бодит текст ангилагч бий болгож байна, одоо!
- Би яавал дээр вэ?

Сургалтын өгөгдөл байхгүй юу?
Гараар бичсэн дүрмүүд

Хэрэв (wheat эсвэл grain) болон (whole эсвэл bread) биш бол

grain гэж категорло

- өгөгдөл дээр гараар болгоомжтой ажиллах хэрэгтэй
  - Цуглуулж байгаа өгөгдөл дээр гараар тааруулах
  - Цаг үрдэг: анги бүрт 2 өдөр

Маш бага хэмжээтэй өгөгдөл байвал?

- Найв Бейсийг ашигла
  - Найв Бейс бол “high-bias” алгоритм (Ng and Jordan 2002 NIPS)
- илүү их тэмдэглэсэн өгөгдөл ол
  - чамд хүмүүс өгөгдөл тэмдэглээд өгөх ухаантай арга ол
- хагас supervised сургалтын аргууд туршаад үз:
  - Bootstrapping, тэмдэглээгүй документууд дээр EM, …

Bootstrapping – бага хэмжээний өгөгдөл ашиглан их хэмжээний өгөгдөл сургадаг арга

Боломжийн хэмжээний өгөгдөл байвал?

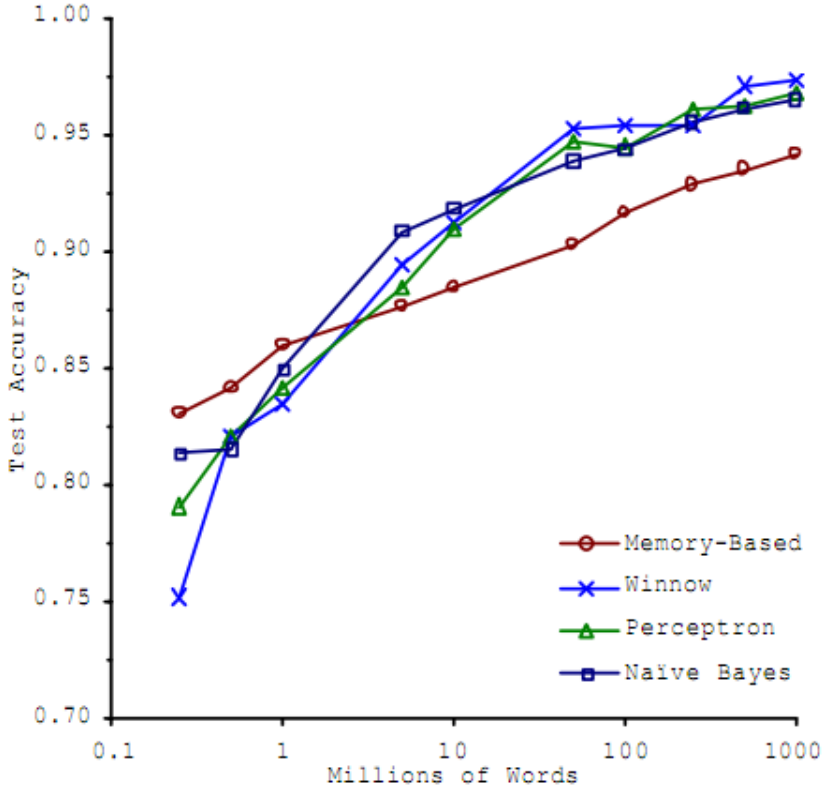
- Бүх ухаалаг ангилагчийн хувьд төгс
  - SVM
  - Regularized Logistic Regression–хялбарчилсан логик регресс
- хэрэглэгчийн оруулсан шийдвэрийн мод ашиглаж болно
  - хэрэглэгчид гараар өгөх дуртай
  - тургэн засах боломжтой

Их хэмжээний өгөгдөл байвал?

- Өндөр accuracy –д хүрэх боломжтой!
- Зардлын хувьд:
  - SVMs (сургалтын хугацаа) эсвэл kNN (тестийн хугацаа) хэтэрхий удаан байна
  - Regularized logistic regression ямар ч байсан арай дээр байж болно
- Иймээс л Найв Бейс эргээд хэрэг болдог!

Өгөгдлийн хэмжээнээс accuracy хамаарах

- Хангалттай өгөгдөлтэй үед
  - Ангилагч чухал биш байх магадлалтай



Brill болон Banko зөв бичгийн алдаа зүгшрүүлэлт дээр

Бодит систем ерөнхийдөө хослуулдаг:

- Автомат ангилал
- тодорхойгүй/хүнд/“шинэ” тохиолдолуудын гар аргаар сонголт авдаг

Утга халилтаас сэргийлэх: log –ийн орон зай

- Олон магадлалыг үржүүлэхэд хөвөгч таслал хальдаг.
- Иймээс log(*xy*) = log(*x*) + log(*y*)
  - магадлалуудыг үржүүлэхийн оронд магадлалуудын логуудын нийлбэр олох нь хялбар.
- нормчлогдоогүй хамгийн өндөр лог магадлалын утгатай анги нь хамгийн магадлалтай хэвээр байна.

$$c_{NB} = \underset{c_j \hat{I}\ C}{\operatorname{argmax}} \log P(c_j) + \underset{i \hat{I}\ positions}{\mathring{a}} \log P(x_i \mid c_j)$$

- Загвар нь одоо зөвхөн жингүүдийн нийлбэрийн хамгийн их утга болно.

Гүйцэтгэлийг хэрхэн сайжруулах вэ?

- Хэрэглээний мужид тохирсон шинж чанар болон жин: бодит гүйцэтгэлд *маш* чухал
- Заримдаа нэр томьёог хураангуйлах хэрэгтэй болдог:
  - Тоон хэмгүүд, химийн томьёо, …
  - Гэвч стемминг хийх ерөнхийдөө тус болдоггүй.
- Хэтрүүлж жигнэх: 2 удаа таарсан мэт тоолох:
  - гарчиг үгс (Cohen & Singer 1996)
  - параграф бүрийн эхний өгүүлбэр (Murata, 1999)
  - Гарчиг үгсийн агуулж байгаа өгүүлбэр дотор (Ko *et al,* 2002)