# СКАМ/СПАМ СЭТГЭГДЭЛ ТАНИГЧ

## Магадлал Статистик, 6-р баг

М.Ариунзаяа, С.Буян-эрдэнэ, О.Нямбаяр, Б.Эрдэнэ-очир

# ЯАГААД ЭНЭ СЭДВИЙГ СОНГОХ БОЛСОН ШАЛТГААН

Нийгмийн сүлжээнд хуурамч мэдээлэл, зээл санал болгосон scam/spam сэтгэгдэл ихэссэн тул хэрэглэгчдийг хамгаалах зорилгоор ийм төрлийн текстийг автоматаар илрүүлдэг ухаалаг системийг хөгжүүлэх шаардлагатай гэж үзээд энэ сэдвийг сонгосон.

# өгөгдөл

| label | Коммент бичсэн хүний нэр | Raw comment | Постны агуулга | Зураг агуулсан эсэх |
|---|---|---|---|---|
| Spam | Corneille Maff | Zeel olgoj baina. Khvv 1,5% | zeel | 0 |
| Spam | Veronique Battuza | Sain baina uu,<br>Bid 2%-iin khüütei, khurdan bögööd naidvartai zeel sanal bolgoj baina. Enekhüü sanalyg ashiglakhyn tuld bidentei kholbogdono uu. | zeel | 0 |
| Ham | Tulgaa Tulgaa | Qpaytei gazraas xudaldan awalt xiine dans edrluu tataj bloxgui | zeel | 0 |
| Spam | Khüütei Medeelel | Khüütei möngö zeelne geree khiine khödöö oron nutag khamaarakhgüi chataar medeelel avaarai ✍️🕊️🕊️ | zeel | 0 |

| Нэрээ нууцалсан эсэх | Монгол нэр эсэх | Цэвэрлэсэн сэтгэгдэл | Голдуу ашигласан үсэг | Кирил, латин биш тэмдэгт ашигласан эсэх | Emoji-ний тоо | Сэтгэгдлийн урт | Email агуулсан эсэх | Link агуулсан эсэх | Утасны дугаар агуулсан эсэх |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | Зээл олгож байна. Хүү 1,5% | Latin | 0 | 0 | 27 | 0 | 0 | 0 |
| 0 | 0 | Сайн байна уу, Бид 2%-ийн хүүтэй, хурдан бөгөөд найдвартай зээл санал болгож байна. Энэхүү саналыг ашиглахын тулд бидэнтэй холбогдоно уу. | Latin | 1 | 0 | 143 | 0 | 0 | 0 |
| 0 | 1 | QPay-тай газраас худалдан авалт хийнэ данс энэ тэр лүү татаж болохгүй | Latin | 0 | 0 | 62 | 0 | 0 | 0 |
| 0 | 1 | Хүүтэй мөнгө зээлнэ, гэрээ хийнэ хөдөө орон нутаг хамаарахгүй чатаар мэдээлэл аваарай. ✍️🕊️🕊️ | Latin | 1 | 3 | 94 | 0 | 0 | 0 |

# АШИГЛАСАН АЛГОРИТМУУД

## Decision Tree

Decision Tree нь өгөгдлөөс хамгийн чухал асуултуудыг олж, тэдгээрийг дарааллуулан “Хэрвээ — Тийм / Үгүй” хэлбэрээр асууж, эцэст нь шийдвэрийг гаргадаг ухаалаг шийдвэр гаргах мод юм.

## Naive Bayes

Naive Bayes нь өгөгдсөн текстийн үг тус бүрийн магадлалыг Spam болон Ham ангиллууд дээр тооцоолж, ямар ангилал хамгийн өндөр магадлалтай вэ гэдгийг Байесийн томъёогоор бодож шийдвэр гаргадаг алгоритм юм.

# PARENT ENTROPY /GAIN

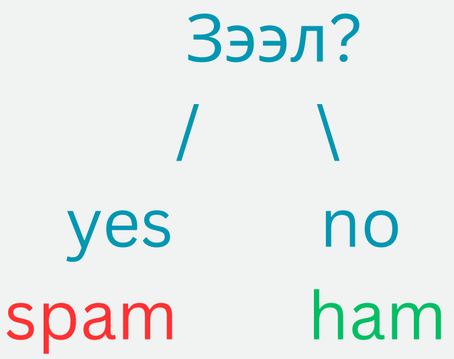
| comment | Зээл | Мөнгө | spam/ham |
|---|---|---|---|
| 1 | yes | yes | spam |
| 2 | yes | no | spam |
| 3 | no | yes | ham |
| 4 | no | no | ham |

1

Entropy=−pspamlog2(pspam)−phamlog2(pham)
pspam=2/4=0.5,  pham=0.5
Entropy(parent)=− 0.5log2*0.5−0.5log2*0.5=1

2

Entropy(зээл=yes)=0
Entropy(зээл=no)=0
Gain=Entropy(parent)−(0.5×0+0.5×0)
Gain(зээл) = 1 - 0 = 1



# БАЙЕСИЙН ТОМЬЁО

$$P(\text{Spam} \mid \text{Comment}) = P(\text{Comment} \mid \text{Spam}) \cdot P(\text{Spam}) / P(\text{Comment})$$

$$P(\text{Ham} \mid \text{Comment}) = P(\text{Comment} \mid \text{Ham}) \cdot P(\text{Ham}) / P(\text{Comment})$$

Spam ангиллын магадлалууд

- $P(\text{Spam}) = 0.4$
- $P(\text{loan} \mid \text{Spam}) = 0.8$
- $P(\text{money} \mid \text{Spam}) = 0.6$

Ham ангиллын магадлалууд

- $P(\text{Ham}) = 0.6$
- $P(\text{loan} \mid \text{Ham}) = 0.1$
- $P(\text{money} \mid \text{Ham}) = 0.05$

$$P(\text{Spam} \mid \text{loan money}) = P(\text{loan} \mid \text{Spam}) \times P(\text{money} \mid \text{Spam}) \times P(\text{Spam}) = 0.8 \times 0.6 \times 0.4 = 0.192$$

$$P(\text{Ham} \mid \text{loan money}) P(\text{loan} \mid \text{Spam}) \times P(\text{money} \mid \text{Spam}) \times P(\text{Spam}) = 0.1 \times 0.05 \times 0.6 = 0.003$$

$0.192 > 0.003 \Rightarrow$ Коммент SPAM гэж ангилна.

# ҮР ДҮНГИЙН ХАРЬЦУУЛАЛТ

**Decision Tree**

Decision Tree алгоритм нь өгөгдлийг ангилахад 62.30% нарийвчлалтай ажилласан бөгөөд “ham” (энгийн сэтгэгдэл)-ийг харьцангуй сайн таньсан ч “spam” сэтгэгдлийг таних чадвар нь харьцангуй сул гарсан байна. Confusion matrix-ээс харахад систем нийт 73 энгийн сэтгэгдлийг зөв ангилсан боловч 45 spam сэтгэгдлийг буруу ангилан ham гэж андуурсан нь spam илрүүлэлтийн гүйцэтгэлд гол нөлөө үзүүлсэн

**Naive Bayes**

Naive Bayes модел нь нийт тестийн өгөгдөл дээр 93.77%-ийн өндөр нарийвчлалтай ажилласан нь энэ алгоритм тухайн асуудалд хамгийн үр дүнтэй ангилаж байгааг харуулж байна. Confusion matrix-ээс харахад, модел ham ангиллыг 143 тохиолдолд зөв, 13 тохиолдолд буруу ангилсан бол spam ангиллыг 113 удаа зөв, ердөө 4 удаа андуурсан байна.

# ДҮГНЭЛТ

Манай хөгжүүлсэн Spam сэтгэгдэл илрүүлэгч жижиг систем нь одоогийн байдлаар зөвхөн зээлтэй холбоотой спам сэтгэгдлийг өндөр нарийвчлалтайгаар илрүүлэх боломжтой . Энэ нь хэрэглэгчдэд өдөр тутмын пост, комментуудаас зээлийн шинжтэй сэжигтэй мэдээллийг хурдан таньж, аюулгүй байдлаа хангахад үр дүнтэй шийдэл болж байна.
Гэсэн хэдий ч бүх төрлийн спам (сурталчилгаа, залилан, фишинг, залилангийн холбоосууд гэх мэт)–ыг өндөр нарийвчлалтай илрүүлэхийн тулд илүү олон төрөл, олон эх сурвалжаас бүрдсэн том хэмжээний сургалтын өгөгдлийн сан шаардлагатай.
Иймээс ирээдүйд системийг дараах чиглэлээр өргөжүүлэх боломжтой:

- Спамын төрөл бүрийн өгөгдлийг өргөн хүрээтэй цуглуулах
- Мэдээлэл боловсруулалтын нарийвчилсан аргачлал сайжруулах
- Илүү хөгжингүй ML/DL загвар ашиглан бүх төрлийн спамыг илрүүлэх чадварыг нэмэгдүүлэх

Эцэст нь, энэхүү төсөл нь анхан шатны түвшинд бодит хэрэглээнд ашиглагдах боломжтой, цаашид хөгжүүлбэл олон төрлийн spam-ийг илрүүлдэг бүрэн хэмжээний ухаалаг систем болох боломжтой гэж дүгнэж байна.

# АШИГЛАСАН МАТЕРИАЛ

- **Decision Tree Learning**

Russell, S., & Norvig, P. (2010). Decision tree learning. In Artificial intelligence: A modern approach (3rd ed., pp. 653–662). Prentice Hall.

- **Naive Bayes Models**

Russell, S., & Norvig, P. (2010). Naive Bayes models. In Artificial intelligence: A modern approach (3rd ed., pp. 758–760). Prentice Hall.

# КОД & НЭЭЛТТЭЙ ӨГӨГДӨЛ

- **Github дээр байршуулсан код:**

https://github.com/nyambayar0118/mgl_facebook_comment_classifier

- **Google sheets дээр байршуулсан өгөгдөл:**

https://docs.google.com/spreadsheets/d/10lRFCELKMqHrkSAQwI2f4jBtAzvPQvkSklUJj84SgLE/edit?usp=sharing

# АНХААРАЛ ХАНДУУЛСАНД БАЯРЛАЛАА