Teoiric na Faisnéise agus an Ghaeilge

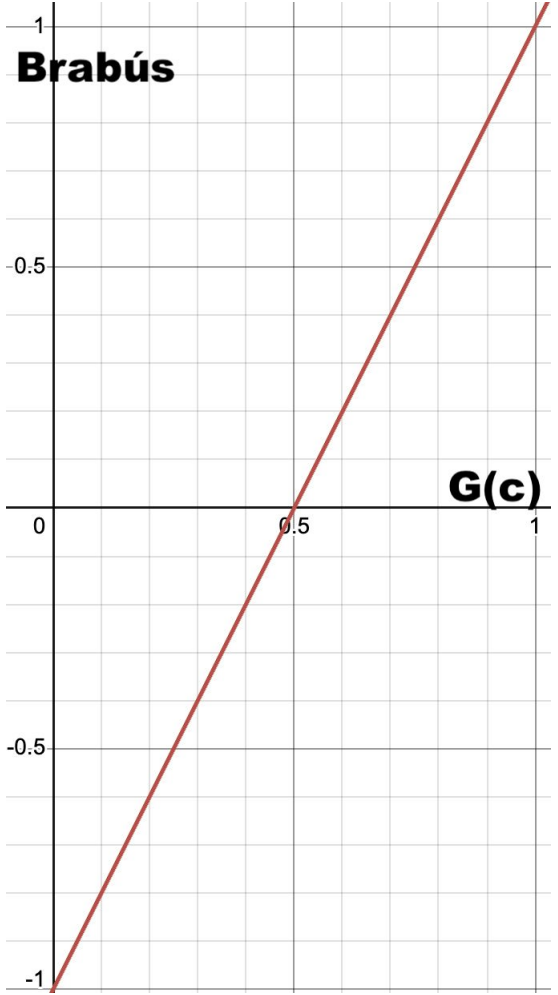
Kevin Scannell
Cadhan Aonair
12 Deireadh Fómhair 2025

Cluiche Cearrbhachais

- Tabharfaidh mé tús abairte duit (m.sh. “Ní bheidh a leithéid _____”)
- Caithfidh tú geall (€1) a chur ar an chéad fhocal eile
- Is féidir leat an t-airgead a “leathnú amach” ar go leor focal éagsúla
- Ansin, babhta eile dírithe ar an chéad fhocal eile, srl srl

Brabús agus caillteanas

- Tugaimis “***c***” ar an bhfocal ***c***eart
- Agus 0 <= G(c) <= 1 ar an n***G***eall a chuir tú air
- An íocaíocht = 2G(c)
- An brabús = 2G(c) - 1
- An caillteanas = 1 - 2G(c)



Cluiche fadtéarmach

- Sprioc: an brabús is mó a fháil *ar an meán, san fhadtéarma* (-1 <= €/geall <= 1)
- Tá dóchúlachtaí éagsúla ar gach focal, ag brath ar an gcomhthéacs
- Sa sampla a bhí againn, “Ní bheidh a leithéid _____”
- Is féidir *meastachán* a dhéanamh ar na dóchúlachtaí ó chorpas
- P(arís) ≃0.46
- P(ann) ≃ 0.20
- P(de) ≃ 0.15
- P(aríst) ≃ 0.05
- ...
- Brabús fadtéarmach ≃ 2 ∑ P(w)G(w) - 1 (∑P(w)=1 agus ∑G(w)=1)

Sampla #2

- “Beidh cóisir againn Dé _____”
- P(Sathairn) ≃ 0.44
- P(Domhnaigh) ≃ 0.19
- P(hAoine) ≃ 0.17
- P(Máirt) ≃ 0.11
- P(Céadaoin) ≃ 0.06
- ...

Sampla #3

- “Thug mé cuairt ar Bhaile Átha _____”
- P(Cliath) ≃ 0.964
- P(Luain) ≃ 0.012
- P(an) ≃ 0.005
- P(Troim) ≃ 0.004
- …
- (“Nuair a bhí mé i gContae na hlarmhí, thug mé cuairt ar Bhaile Átha _____”?)

Sampla #4

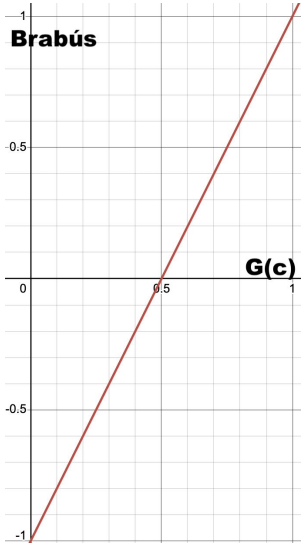
- “ar chor ar _____”
- P(bith) ≃ 0.998
- 0 < P(bhith) < 0.001
- 0 < P(both) < 0.001
- 0 < P(bit) < 0.001
- 0 < P(bioth) < 0.001

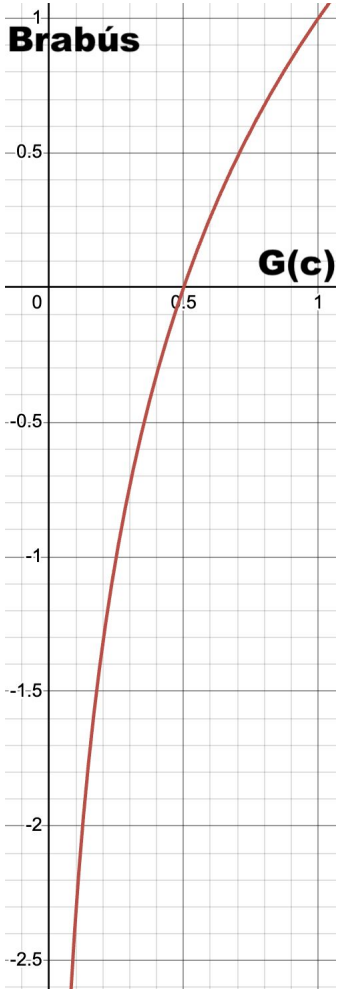
Sampla #5 (níos tipiciúla)

- “Tá sé _____”
- P(ag) ≃ 0.078
- P(i) ≃ 0.064
- P(ar) ≃ 0.043
- P(sin) ≃ 0.035
- P(ráite) ≃ 0.034
- P(in) ≃ 0.024
- … na mílte mílte eile, cosúil le crannchur!

Cluiche níos spéisiúla!

- An cluiche céanna, le hathrú amháin
- Íocaíocht = $2 + \log_2 G(c)$
- Brabús = $1 + \log_2 G(c)$
- Caillteanas = $- 1 - \log_2 G(c)$
- Brabús fadtéarmach $\approx 1 + \sum P(w)\log_2 G(w)$





Cad é an straitéis is fearr?

- Abair gur féidir leat meastachán cruinn a dhéanamh ar gach P(w)
- Leis sin, is é an straitéis is fearr G(w) = P(w)
- Bíonn P(w)>0 i gcónaí, agus mar sin G(w)>0 i gcónaí (ach an-bheag go minic)
- An brabús fadtéarmach = $1 + \sum P(w)\log_2 G(w) = 1 + \sum P(w)\log_2 P(w) = 1 - H(P)$
- Tugtar “eantrópacht Shannon” ar H(P) (tar éis Claude Shannon)
- H(P) = an méid ***faisnéise*** a fhaightear ó léiriú an chéad fhocal eile, ar an meán
- De réir na meastacháin is fearr tá H(P) idir 4.0 agus 5.0 (giotáin/focal)
- => Caillfidh tú mórán airgid ag imirt leis na rialacha nua!!

Samhaltú Teanga

- Cad is LLM ann, nó “ollsamhail teanga”?
- Ríomhchlár atá traenáilte chun an cluiche seo a imirt — sin an méid!
- Tá *mórán* eolais ionchódaithe i samhail teanga (chuile rud i samhail “fhoirfe”?)
- Comhréir, gramadach, nathanna cainte, eolas faoin saol, …
- Úsáideann na samhlacha is fearr **líonraí néaracha**
- Déantar traenáil chun an caillteanas $(- 1 - \log_2 G(c))$ a íoslaghdú

Teicneolaíocht Teanga

- Úsáidtear samhlacha teanga i mbeagnach gach teicneolaíocht teanga
- Ríomhaistriúchán
- Aithint cainte
- Botanna comhrá (ChatGPT, Gemini, etc.)
- Samhail teanga níos fearr => uirlis níos fearr, go ginearálta

Seiceálaí Gramadaí

- Anois, cluiche atá i bhfad níos éasca
- An leagan amach céanna, tús abairte, m.sh. “cheap mé go raibh cuma _____”
- Arís, caithfidh geall a chur ar an chéad fhocal eile
- Ach anois, tugaim leid láidir duit!
- *Is é “breá” nó “bhreá” an chéad fhocal eile*
- Is féidir €1 a chur ar cheann amháin nó an ceann eile, nó é a roinnt eatarthu
- NB: Is é $1 + \log_2 G(c)$ an brabús arís!!
- “Níos éasca” == sa chás seo, is féidir linn airgead a ghnóthú!

Teoiric na faisnéise

- .i. Tá eantrópacht na n-athruithe tosaigh sa Ghaeilge an-bheag
- .i. Ní iompraíonn siad mórán eolais/faisnéise
- Dearcadh eile: dá scriosfainn na hathruithe tosaigh, bheadh cainteoir líofa in ann iad a chur ar ais i mbeagnach chuile chás:

Deirtear go iompraíodh sí gunnaí ina carr, iad faoi ceilt i mála plúir.

Ní raibh Gaoth Dobhair ann mar ainm dúiche ná paróiste ar tús, ach mar ainm ar an gaoth / abhainn ónar baisteadh an ceantar, an cainéal nó an inbhear farraige idir an paróiste agus na Rosa, ar a tugtar an Gaoth go dtí an lá inniú, agus an abhainn.

Teoiric na faisnéise

- .i. Tá eantrópacht na n-athruithe tosaigh sa Ghaeilge an-bheag
- .i. Ní iompraíonn siad mórán eolais/faisnéise
- Dearcadh eile: dá scriosfainn na hathruithe tosaigh, bheadh cainteoir líofa in ann iad a chur ar ais i mbeagnach chuile chás:

Deirtear go n-iompraíodh sí gunnaí ina carr, iad faoi cheilt i mála plúir.

Ní raibh Gaoth Dobhair ann mar ainm dúiche ná paróiste ar dtús, ach mar ainm ar an ghaoth / abhainn ónar baisteadh an ceantar, an cainéal nó an t-inbhear farraige idir an pharóiste agus na Rosa, ar a dtugtar an Gaoth go dtí an lá inniú, agus an abhainn.

Torthaí

- Eantrópacht: níl ach **0.07** giotán/focal sna hathruithe tosaigh
- Tástáil le 10000 focal: P>0.5 don athrú tosaigh ceart 98.63% den am (€€€!)
- Tá na 137 “botún” freagrach as 77% den eantrópacht (caillteanas)
- 61 as 137: botúin ghramadaí sa chorpas tástála!!
- 23 as 137: leaganacha cearta nach gcloíonn leis an gcaighdeán (m.sh. *dhom*)
- Baineann 16 le haidiacht shealbhach sa tríú pearsa (*a*, *ina, faoina, …*)
- Baineann 9 gcinn le cúrsaí canúna (séimhiú vs. urú sa tuiseal tabharthach)

Go raibh maith agaibh!

- https://cadhan.com/
- https://github.com/kscanne/