

kapitola 15

Jak stroj uhodne další slovo?

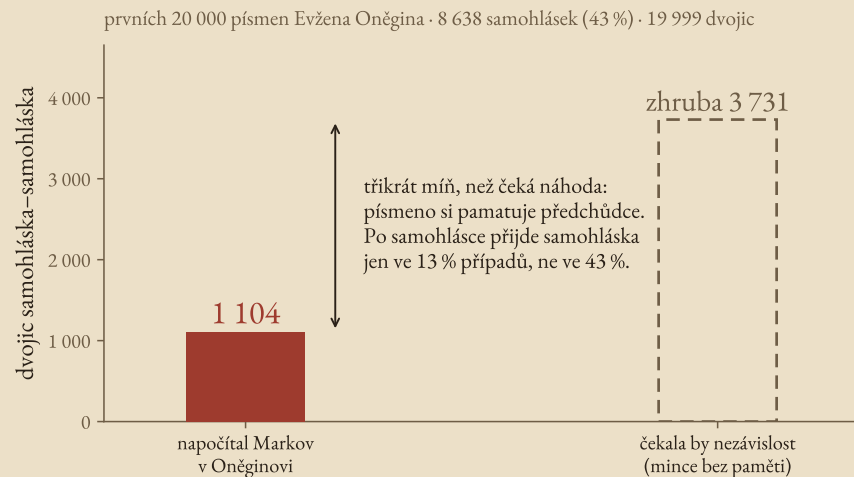
Matematik, který počítal Puškina

1913, Petrohrad. Markov vzal 20 000 písmen *Evžena Oněgina* a každé hodil do jedné ze dvou přihrádek: **samohláska**, nebo souhláska.

Pak týdny počítal ručně. A ukázal, že písmeno **si pamatuje svého předchůdce** — verše nejsou loterie.

Nikdo modelu neřekl jediné pravidlo gramatiky. Jen počítal.

Písmeno má paměť



zdroj: A. A. Markov, 1913; Hayes, American Scientist 101(2), 2013

Náhoda by čekala 3 731 dvojic. Markov jich napočítal 1 104. Tříkrát méně.

Dvě čísla = jazykový model

- Po samohlásce přijde samohláska jen v 13 %, po souhlásce v 66 %.
- Ta dvě čísla už umějí to hlavní: **predikovat** další písmeno.
- Statistika textu nese znalost jazyka — a **kdo počítá, ten ji vytěží**.
- Tomuhle stroji říkáme **Markovův řetěz**: každý článek visí jen na posledním.

Hra, kterou hraješ celý život

1951: Shannonova žena Betty hádá písmeno po písmenu, co v knize následuje. Neměřil ji — měřil **angličtinu v její hlavě**.

V hluku hospody, nad rukopisem lékaře, v křížovce — tvůj mozek **předpovídá pokračování** z jeho statistiky.

Přesně tuhle hru hraje při tréninku i velký jazykový model.

Celý stroj u piva: čtyři pojmy

- **Token** — kousek textu mezi písmenem a slovem.
- **Embedding** — poloha tokenu na mapě významů.
- **Pozornost** — slovo se zeptá ostatních, co je důležité.
- **Teplota** — knoflík mezi úředníkem a básníkem.

Pracovní stůl, ne knihovna

Kontextové okno je **stůl, ne police**: co na něm neleží, pro model neexistuje.

A pozor — v řetězu tokenů **chybí krok „ověř, jestli je to pravda,,**. Model dokonale umí **znít jako pokračování**. Plynulost je kvalita mapy textů, ne mapy světa.

„Jak bych poznal, že se pletu?„

Zahraj si Shannonovu hru dvakrát: na Máchově *Máji* a na svém včerejším e-mailu.

Vyjde-li perplexita různá, víš z první ruky, co znamená „záleží, na čem byl trénovaný“.