人工智能在猎豹内容产品的应用

闵可锐

News Republic介绍

面向海外的个性化内容推荐系统

- 数千正版媒体源（BBC, CNN, AP...）
- 数千万的下载安装量
- Google Play新闻杂志类Top 10

News Republic介绍

每日新增超过10万篇内容

- 分词
- 词性分析
- 命名实体分析
- 提取关键词
- 情感分析
- . . .

为何需要语义分析

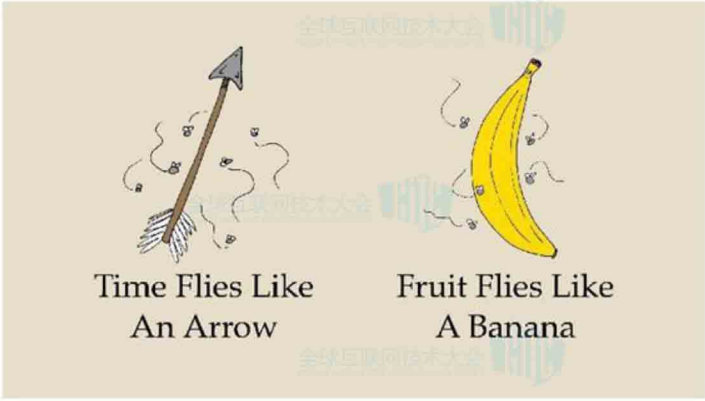
通过对符号字符串的解析，回答一篇文章在讲什么（Meaning），寻找与之相关的文章和潜在感兴趣的用户

理解文章 $\Rightarrow$ 理解用户

为何需要语义分析

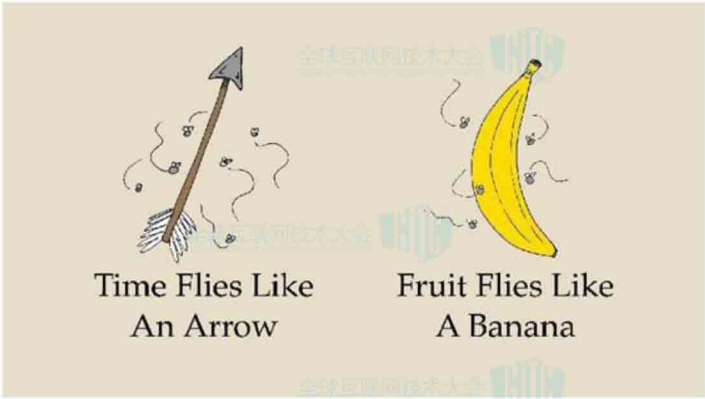
传统基于关键词理解：TF*IDF，Okapi BM25

二义性：从词语到句子，从语法到语义



为何需要语义分析

传统基于关键词理解：TF*IDF，Okapi BM25
二义性：从词语到句子，从语法到语义



Time: 时间/计时/时代周刊? Flies: 飞行/苍蝇? Like：像/喜欢?

向量表示

从词向量到文章向量

- 词向量：$w \in \mathbb{R}^d, w \in V$
- 文章向量：$c \in \mathbb{R}^D, c \in \mathcal{D}$

BOW表达

- $c = \sum_i w_i$，丢失词序信息

简单拼接

- $c = \text{concat}(w_1 w_2 \cdots w_n), D = dn$
- 保留词序，向量空间大小不定

向量表示目标

映射为固定空间大小，方便使用

对词序敏感

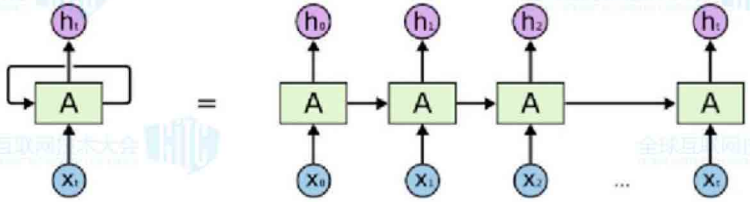
能够 无标注 数据进行训练

向量表示目标

映射为固定空间大小，方便使用
对词序敏感

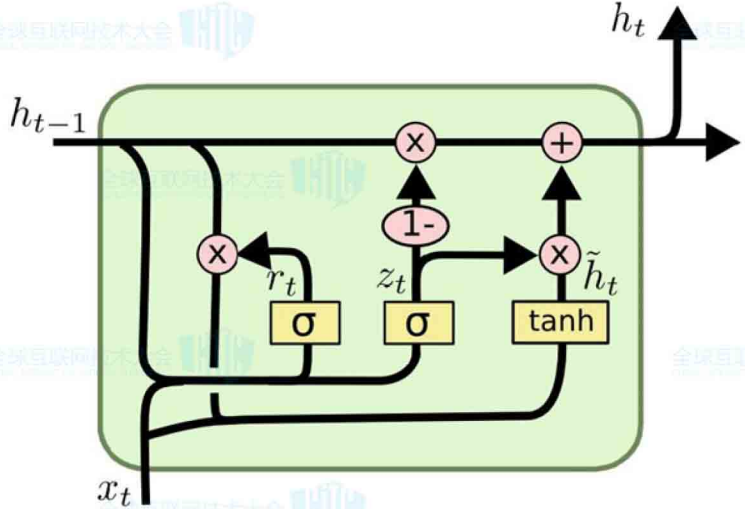
$\Longrightarrow$ **Recurrent Neural Network**

$$h_t = f(h_{t-1}, x_t).$$



GRU

实际我们采用 `Gated Recurrent Unit` 进行实现



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Chung, Junyoung, etc. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling", 2014.

无标注训练

映射为固定空间大小，方便使用

对词序敏感

能够 `无标注` 数据进行训练

```
Title: Microsoft released two brand-new laptops this
week, both killer Apple rivals

Date: 2017-06-17

Publisher: Business Insider

Content: This week, Microsoft released not one, but two new laptops.

First, you've got the Surface Laptop, an all-new device that's aimed at
students. It also serves as the showcase device for Windows 10 S...
```

无标注训练

采用 `Gated Recurrent Unit` 实现 `Encoder-Decoder` 模型

输入：新闻内容（$x_1 \cdots x_n$）

输出：新闻标题（$t_1 \cdots t_l$）

$$\begin{aligned} &\text{Encoder}(x_1 \cdots x_n) \rightarrow h_n \in \mathbb{R}^D, \\ &\text{Decoder}(h_n) \rightarrow t_1 \cdots t_l. \end{aligned}$$

h_n作为我们的语义向量（Why?）

经过实验，我们采用512维向量（$D = 512$）训练结果

一些例子

Verizon Among Leading Bidders for Yahoo's Core Business: Bloomberg

```
1. Verizon Is Said to Be Near a Deal to Acquire Yahoo
2. UPDATE 1-Yahoo to be named Altaba, Mayer to leave
board after Verizon deal
3. Verizon Communications : Yahoo's Marissa Mayer to
resign from board after Verizon deal closes
```

SWAT team called to standoff incident in College Park

```
1. Woman shot in car near East Nashville apartments
2. Uber Driver Strikes Man 'Clearly Outside of
Crosswalk:' San Jose Police Related
3. Man shot after confronting another driver over tailgating
```

效率优化

对于任意两个语义向量距离度量

$$\|h_i - h_j\|_2^2 = \sum_{k=1}^{D}(h_{ik} - h_{jk})^2$$

上千次浮点运算.

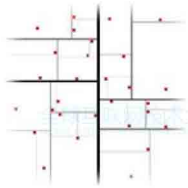
关注近邻查找而不是向量本身.

$$h^* = \mathcal{Q}(h_q) = \min_{h \in \mathcal{H}} \|h - h_q\|_2$$

近似解方案

KD-Tree

- 划分空间索引，搜索剪枝
- 精确查找时间复杂度：$O(\log n) \rightarrow O(n)$
- 近似查找时间复杂度：$O\left(\frac{1}{\epsilon^D} \log n\right)$
- 维数灾难



近似解方案

Locality-sensitive Hashing

对任意$p, q \in \mathcal{M}$, 均匀采样$h \in \mathcal{F}$满足

- 如$d(p,q) \leq R$, $\mathbb{P}(h(p) = h(q)) \geq P$;
- 如$d(p,q) > cR$, $\mathbb{P}(h(p) = h(q)) \leq P - \epsilon$.

存在$\mathcal{F}$，$\rho < 1/c$，使得

- 近似查找时间复杂度：$O(Dn^{\rho})$
- 索引时间复杂度：$O(Dn^{1+\rho})$

Learning to Hash

LSH所采用的数据划分方式与具体数据无关（data-independent）
对于具体数据集，可以根据其分布得到更好的划分

目标：

学习映射f，对于$h_i \in \mathcal{H}$，对应的语义指纹$b_i = f(h_i)$满足：

- 二值化：$b_i \in \{-1, 1\}^d$,
- 近似保距：$\forall i, j$ Hamming$(b_i, b_j) \approx c \cdot \|h_i - h_j\|_2$

Learning to Hash

$$f: \mathbb{R}^D \mapsto \mathbb{R}^d.$$

e.g. $[3.16, 1.05, \cdots, -1.37] \mapsto 101101.$

优势：

- 时间：相似度比较由浮点运算变为整数XOR运算
- 空间：由D个实数变为了$\lceil d/32 \rceil$个int32整数

DeepBit

DeepBit采用深度学习建模$f(\cdot, W)$, W为网络参数：

$$y_i = f(h_i, W), b_i = \text{sign}(y_i).$$

对于一个Mini-batch，采用如下的损失函数：

$$\text{Loss} = L_{\text{quant}} + \alpha L_{\text{even}}$$

- $L_{\text{quant}} = \|y_i - b_i\|_2$
- $L_{\text{even}} = \sum_k \left(\sum_{i \in \mathcal{B}} b_{ik}\right)^2$

Kevin Lin, etc. Learning Compact Binary Descriptors with Unsupervised Deep Neural Networks, CVPR 2016.

DeepBit

DeepBit采用深度学习建模$f(\cdot, W)$, W为网络参数：

$$y_i = f(h_i, W), b_i = \text{sign}(y_i).$$

Algorithm 1: DeepBit

Input: Training set $X = [x_1, x_2, ..., x_n]$
Output: A set of non-linear projection parameters $\mathcal{W}$

Step 1 (Initialization):
Initialize $\mathcal{W}$ with pre-trained weights from ImageNet;

Step 2 (Optimization):
while $iter < max\_iter$ **do**
  Fix $\mathcal{W}$ update b_n using (1);
  while $iter1 < max\_iter1$ **do**
    Fix b_n update $\mathcal{W}$ by minimizing the sum of (5) and (6);
  Fix $\mathcal{W}$ update b_n using (1);
  while $iter2 < max\_iter2$ **do**
    Fix b_n update $\mathcal{W}$ using (8);

return $\mathcal{W}$;

DeepBit

上述算法采用的是 `Alternating Minimization` 的方式，交互更新W与b_i：

- 收敛速度较慢
- 更容易陷入次优（sub-optimal）的局部最优解

为什么不采用 `端到端` 的方式进行优化?

端到端优化

```
def train(h, W, alpha):
    # forward step
    y = f(h, W)
    b = sign(y)

    L = L_quant(y, b) + alpha*L_even(b)
    dW = backprop(L)

    W = W - lr*dW # SGD step
```

是否可行?

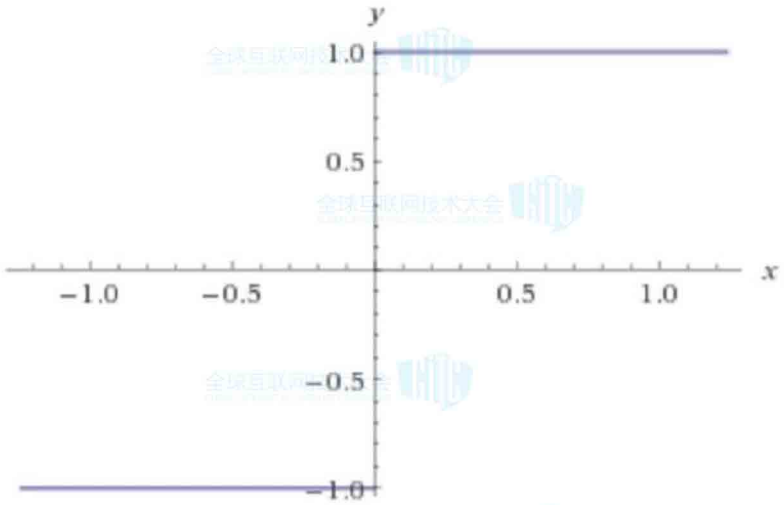
No!

```
def train(h, W, alpha):
    # forward step
    y = f(h, W)
    b = sign(y) # <================ (*)

    L = L_quant(y, b) + alpha*L_even(b)
    dW = backprop(L)

    W = W - lr*dW # SGD step
```

Sign函数



导数*几乎处处*为零！

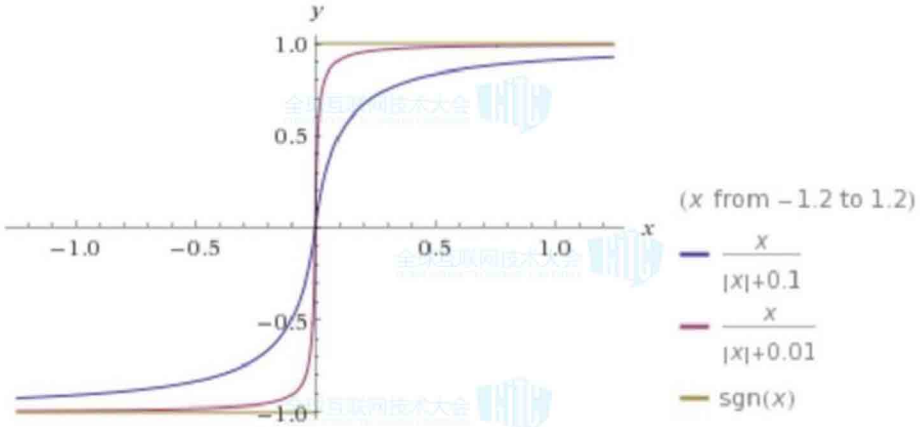
在端到端更新中，没有可被反向传播的梯度

近似Sign函数

可以采用soft-sign函数对sign函数进行近似

$$\text{soft-sign}(x, \epsilon) = \frac{x}{\epsilon + |x|},$$

$$\text{sign}(x) = \lim_{\epsilon \to 0} \text{soft-sign}(x, \epsilon).$$



近似准确率

$$\operatorname{sign}(x) = \lim_{\epsilon \to 0} \text{soft-sign}(x, \epsilon).$$

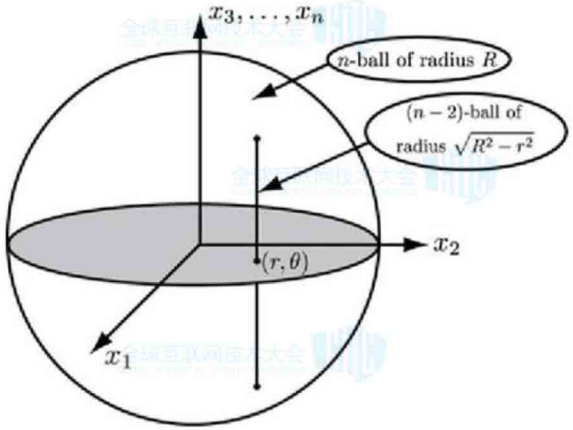
对于x，近似程度取决于$\epsilon / |x|$的值.

为避免由于$y = f(x, W)$过小带来的，引入 倒数正则项 ：

$$L^2_{\text{inverse}} = \left\| \frac{1}{y} \right\|_2^2 = \sum_{i=1}^{d} \frac{1}{y_i^2}.$$

非均匀采样

近似保距：$\forall i, j \ \text{Hamming}(b_i, b_j) \approx c \cdot \|h_i - h_j\|_2$



$$L_{\text{dist}} = L_{\text{KNN}} + L_{\text{rand}}.$$

L_{KNN}：K近邻近似误差，L_{rand}随机点近似误差.

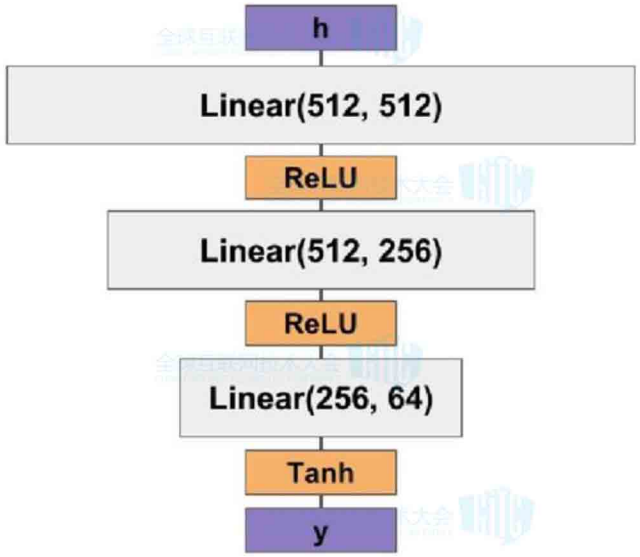
Put together

```
def train(h, W, alpha):
    # forward step
    y = f(h, W)
    b = soft-sign(y) # approximate sign function

    L = L_dist(b, h) + alpha*L_inverse(y)
    dW = backprop(L)

    W = W - lr*dW # SGD step
```

定义 f 网络

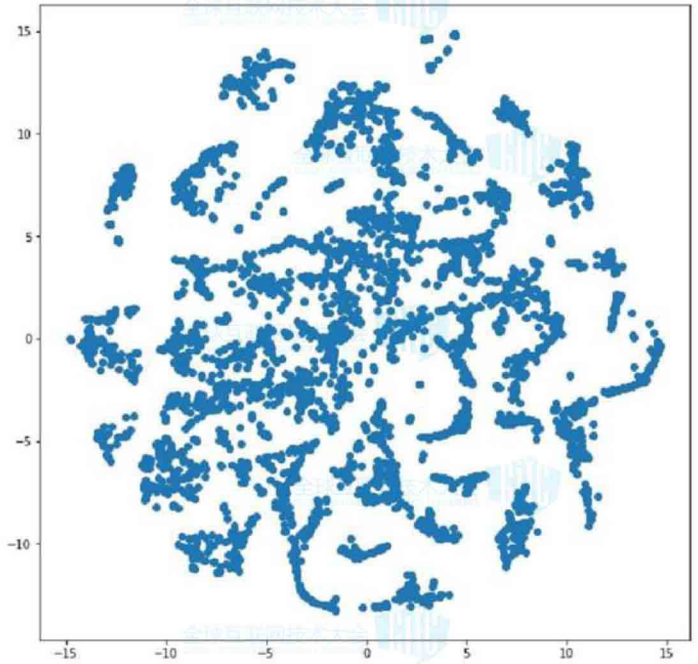


简单三层网络，总参数～1.6MB

实验结果

Batch-size=32完成50k次SGD迭代

固定$\epsilon = 0.01$，99%的$b_{ij} \in [-1, -0.99] \cup [0.99, 1.0]$.



语义指纹碰撞

枪击/谋杀，被判处死刑

```
1. DA's office to seek death penalty in case against
   man accused of killing Kayla Gomez-Orozco
2. Whitehall shooting victim shot himself, lied to
   police about drive-by
3. Jury continues to mull attempted-murder charge in
   Clinton Township shooting
4. Inmate bites caseworker, punches guards at Nebraska
prison
```

语义指纹碰撞

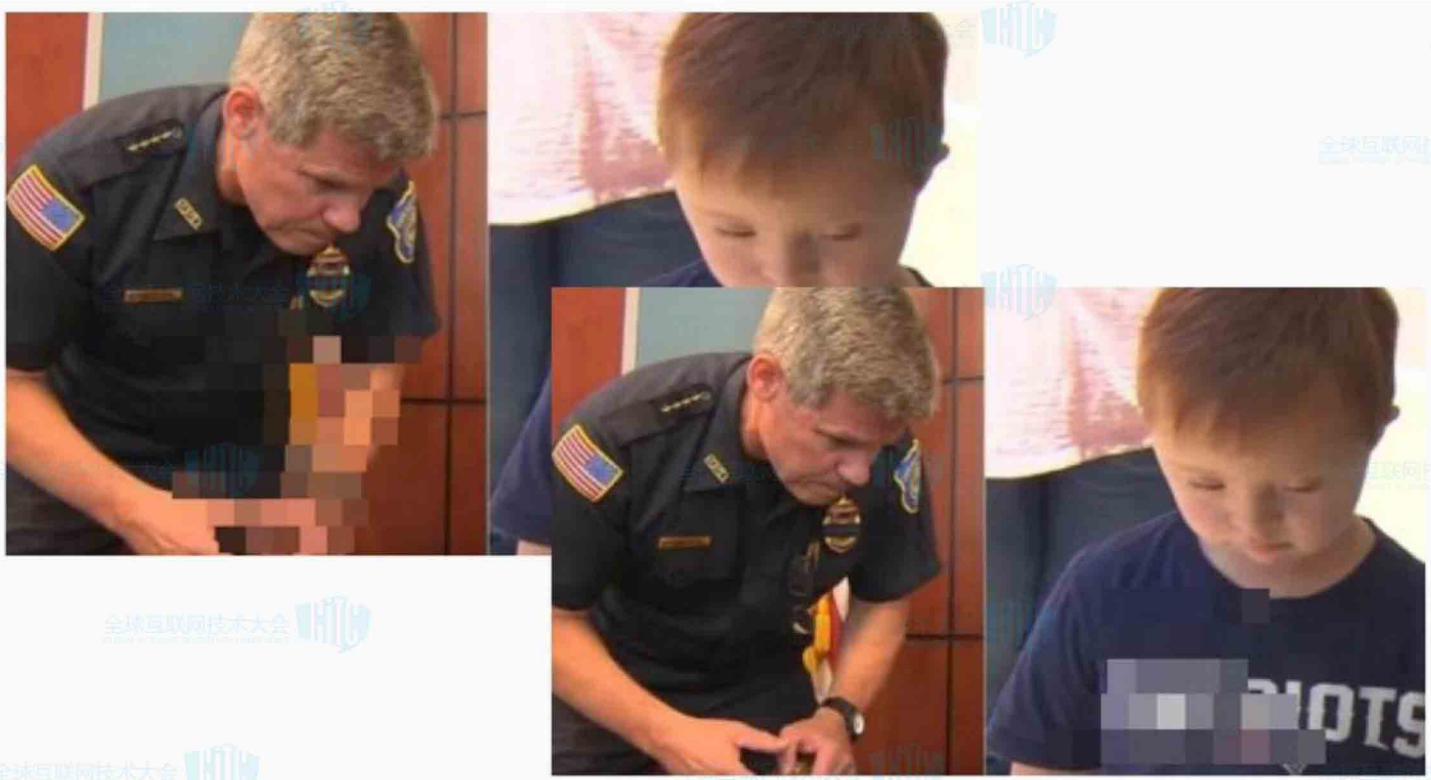
朝鲜导弹试验

1. US see increased activity by Chinese bomber aircraft
2. N. Korea Warns of ‘Super-Mighty Preemptive Strike’ as U.S. Plans Next Move
3. Donald Trump: If China wanted to end North Korean aggression, it could
4. U.S. sets up missile defence system in South Korea as North flexes muscles

深度学习图像理解应用

- 性感/色情图片识别
- LOGO/文本识别
- 不适图片识别
- 马赛克图片识别
- . . .

生成马赛克图片




HSI


NAVY

识别出的文字图片


Všechno je super, když nejsi losers,
máš vlastní plány a hodně řešení


MINECRAFT
POCKET EDITION
hola
0.11.0
hello
LETRAS DE COLORES
TEXTO
hola

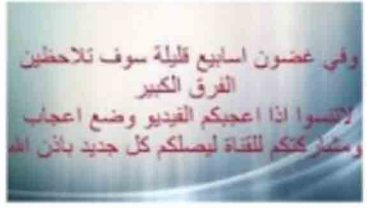

sadomizzo le aspettative
schiavizzo questa folla
che elettrizzo con ste rime,
perso in questa storia di
qui non so le rime,
verso sangue incandescente
che brucia le puntine.


Милосердного
СУРА "ИБРАХИМ"


nessuno
divide

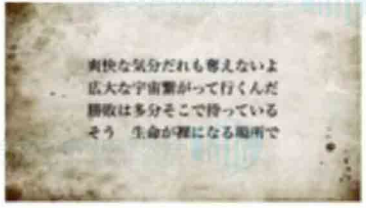

RÁDA
SNĚHULÁKY
STAVÍŠ
KARAOKE
FROZEN

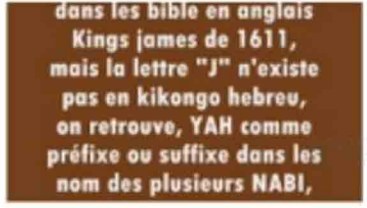

ma sei stato
perdonato per il tuo
silenzio
but you've been forgiven for
your silence


وفي غضون اسابيع قليلة سوف تلاحظين
الفرق الكبير
لاتنسوا اذا اعجبكم الفيديو وضع اعجاب
ومشاركتكم للقناة ليصلكم كل جديد باذن الله


FogVisitor


IDE STUDIO


爽快な気分だれも奪えないよ
広大な宇宙繋がって行くんだ
勝敗は多分そこで待っている
そう 生命が輝になる場所で


dans les bible en anglais
Kings james de 1611,
mais la lettre "J" n'existe
pas en kikongo hebreu,
on retrouve, YAH comme
préfixe ou suffixe dans les
nom des plusieurs NABI,

Thank you! 👍

minkerui@cmcm.com