work & conclusions regarding vectors of skip-gram algorithm

1. first, for computational reasons, we start by normalizing all result vectors of the skip-gram algorithm - $\forall_{v\in\Omega}\, v = \frac{v}{||v||}$.

2. let $d : \Omega^2 \to [0,2]$ be a random variable s.t Ω is a vector space of normalized skip-gram vectors and d measures the euclidean distance between 2 random vectors $d(a,b) \in [0,2]$.

3. we can think of the our vetors with size n as vectors on S^n. and their eucludean distance is at most 2.

4. our goal is to asses the euclidean distance distribution of 2 random vectors from Ω. we denote $d_{a,b} := d(a,b)$.
   let $(a,b) \in \Omega^2$ be a tuple of 2 random vectors. hence by definition:

   $$d_{a,b}^2 = ||a-b||^2 =< a-b, a-b >=< a,a > -2\cdot < a,b > + < b,b >$$

   then follows:

   $$d_{a,b}^2 = ||a||^2 + ||b||^2 - 2\cdot < a,b >= 2 - 2\cdot < a,b >$$

5. using the formula

   $$< a,b >= ||a|| \cdot ||b|| \cdot cos(\theta) \to < a,b >= cos(\theta)$$

   then we get

   $$d_{a,b}^2 = 2 - 2 \cdot cos(\theta)$$

   therefore

   $$d_{a,b} = \sqrt{2 - 2 \cdot cos(\theta)}$$

6. to simplify $d_{a,b}$ even further,

   $$cos(\theta) = 2 \cdot cos^2(\frac{\theta}{2}) - 1$$

   hence

   $$d_{a,b} = \sqrt{2 - 2 \cdot cos(\theta)} = \sqrt{2 - 2 \cdot (2 \cdot cos^2(\frac{\theta}{2}) - 1)}$$

   $$d_{a,b} = \sqrt{4 - 4 \cdot cos^2(\frac{\theta}{2})} = 2 \cdot sin(\frac{\theta}{2})$$

7. finally we can specify a relation between $d_{a,b}$ and another random variable $\theta : \Omega^2 \to [0, 2\pi]$ (denoted by $\theta_{a,b}$):

   $$\frac{d_{a,b}}{2} = sin(\frac{\theta_{a,b}}{2}) \Longleftrightarrow arcsin(\frac{d_{a,b}}{2}) = \frac{\theta_{a,b}}{2}$$

8. **conclusion:** $0 \leq \frac{\theta_{a,b}}{2} \leq \frac{\pi}{2}$

   proof:

   we know that $0 \leq d_{a,b} \leq 2$ therefore $0 \leq \frac{d_{a,b}}{2} \leq 1 \rightarrow 0 \leq sin(\frac{\theta_{a,b}}{2}) \leq 1 \rightarrow 0 \leq \frac{\theta_{a,b}}{2} \leq \frac{\pi}{2}$.

9. **claim:**

   let $\theta_{a,b} : \Omega^2 \rightarrow \mathbb{R}$ be random variable that messures the angle between 2 random vectors $a, b \in S^n$.

10. our method for determining $d_{a,b}$ distribution is to evaluate $\frac{1}{2} \cdot d_{a,b}$ and $\frac{1}{2} \cdot \theta_{a,b}$ distributions from the data. then we will connect the 2 random variables according to the identity above to confirm the results.

11. we will start by calculating

$$skewness := \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1) \cdot \left(\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^3}$$

   and

$$kurtosis := \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1) \cdot \left(\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^4}$$

   of $\frac{1}{2} \cdot d_{a,b}$ and $\frac{1}{2} \cdot \theta_{a,b}$ distribution graphs.

   (a) $\frac{d_{a,b}}{2}$ skewness: -0.60156

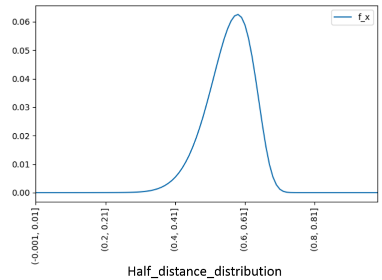
   (b) $\frac{\theta_{a,b}}{2}$ skewness: -0.454

   (c) $\frac{d_{a,b}}{2}$ kurtosis: 3.4324

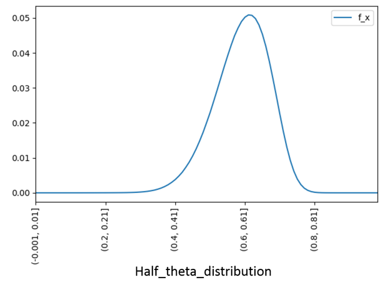
   (d) $\frac{\theta_{a,b}}{2}$ kurtosis: 3.166

   (e) **Data**

   notice that for $0 \leq \frac{d_{a,b}}{2} \leq 1$ it's distribution graph:



   hence, from $0 \leq \frac{d_{a,b}}{2} \leq 1 \rightarrow 0 \leq sin(\frac{\theta_{a,b}}{2}) \leq 1 \rightarrow 0 \leq \frac{\theta_{a,b}}{2} \leq \frac{\pi}{2}$ it's distribution graph:



from The Skewness-Kurtosis the values are close to a normal distribution (skewness: 0, kurtosis: 3) and from the distribution graphs we will fit $\frac{\theta_{a,b}}{2} \sim SN(\xi, \omega^2, \alpha)$ (skewd normal distribution).

12. **a short summary of a skewd normal distribution****

13. **claim:**
let X, Y be 2 random variables then if$Y = sin(X) then Y \sim F_X(arcsin(t))$.
proof:
$F_Y(t) = \mathbb{P}(Y \leq t) = \mathbb{P}(sin(X) \leq t) = \mathbb{P}(X \leq arcsin(t)) = F_X(arcsin(t))$

14. **conclusion:**
$if\ X\ random\ variable\ and\ Y = sin(X)\ then\ F_X(t) = F_Y(u)\ (0 \leq t \leq 1,\ 0 \leq u \leq \frac{\pi}{2})$

15. **claim** - $\mathbb{E}[sin(X)] \approx sin(\mu)$ and $Var[sin(X)] = \sigma^2 \cdot cos^2(\mu)$ moreover:

$$if\ X \sim N(\mu, \sigma^2)\ then\ sin(X)\ can\ be\ approximated\ around\ 0$$

$$sin(X) \approx N(sin(\mu),\ \sigma^2 \cdot cos^2(\mu))$$

proof:
first, let's write

$$sin(X) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + ... + ...$$

as a taylor series. and

$$cos(X) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - .... + ...$$

now, notice that as x≈ 0 hence $sin(x) \approx x$ and $cos(x) \approx 1$.
therefore define $Z = X - \mu \rightarrow X = Z + \mu$ where $Z \sim N(0, \sigma^2)$.
then

$$sin(X) = sin(Z + \mu) = sin(Z)cos(\mu) + cos(Z)sin(\mu)$$

from the taylor series we can deduce:

$$sin(X) \approx Zcos(\mu) + sin(\mu)$$

then we get

$$sin(X) \approx N(sin(\mu),\ \sigma^2 \cdot cos^2(\mu))$$

16. coclusions from the data:

    (a) from the data we assume

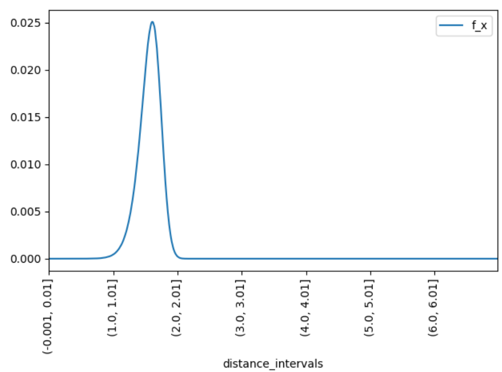
    $$\frac{\theta_{a,b}}{2} \sim N(\mu,\, \sigma^2)$$

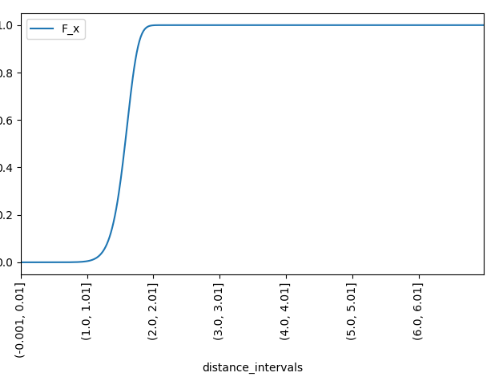
    for

    $$\mu = 0.5959$$

    and

    $$\sigma^2 = 0.006144$$





    (b) from the theory we develop:

    $$\frac{d_{a,b}}{2} \approx N(sin(\mu),\, \sigma^2 \cdot cos^2(\mu)) = N(0.561254, 0.004209)$$

    (c) from the data we get $\frac{d_{a,b}}{2} \approx N(0.559563, 0.004288)$, the result seems fit to our approximation.

17. notice that $\frac{d_{a,b}}{2} \sim N(\mu_1, \sigma_1^2)$ for $\mu_1 \approx 0.56, \sigma_1^2 \approx 0.0042$ meaning $\frac{1}{4} \cdot d_{a,b}^2 \sim \chi_1^2$.

<u>**claim:**</u>

$$if\ for\ all\ v, w \in \Omega\ the\ i'th\ components\ v_i, w_i\ are\ i.i.d\ r.v's\ then$$
$$\forall_v var(v_i) = \frac{1}{2n} \cdot E[d_{a,b}^2]\ hence,\ var(v_i) = \frac{1}{n}.$$

proof:
- notice that

$$E[d_{a,b}^2] = E[\sum_{i=0}^{n}(a_i - b_i)^2] = \sum_{i=0}^{n} E[(a_i - b_i)^2] = n \cdot E[(a_i - b_i)^2] =$$

$$E[d_{a,b}^2] = n \cdot E[a_i^2 - 2a_i b_i + b_i^2] = n \cdot (2E[X_i^2] - 2E[a_i b_i]) =$$

$$\frac{1}{8n} \cdot E[d_{a,b}^2] = E[X_i^2] - E[a_i]E[b_i] = var(X_i)$$

from last claim we know that $d_{a,b}^2 \sim \chi_1^2$ therfore

$$var(X_i) = \frac{1}{n}$$

18. one more interesting result, will be to calculate covariance matrix from the data and estimate if indeed $var(X_i) \approx \frac{1}{100} = 1\%$ for all i. and as we calculate the mean of the variance of all r.v (code: np.diagonal(np.cov(random_sampled_matrix_normalized))) we get a very close result $0.00997 \approx 0.01$. messuring the expectation variance of the vectors minus the expectation vector also get the result $0.010004195023457877 \approx 0.01$.

19. it's only natural trying to figure out how to calculate $E[X_i]$. from the data it seems as $E[X_i] \approx 0.02$.