

بابهتی 4: لهکردنی تۆکن خولی سه‌رهیللی زمانه‌وانیی کۆمپیوتهری

سینا ئه‌حمه‌دی

sinaahmadi.github.io/KurdishCL

1 له تکردنی تۆکن

2 له تکردنی تۆکن بۆ کوردی



لهتکردنی تۆکن



چۆن دهق له سهركۆمپیوتهر نیشان دهدهین؟

له زۆر بابتهی پرۆسهس کردنی زماندا پیویستمان به داشکاندنی دهق (segmentation) بو ده رهینانی
پیکهاته زمانهوانیهکان یان ئاسانکاری تهکنیکی ههیه.



چۆن دەق لە سەر کۆمپیوتەر نیشان دەدەین؟

لە زۆر بابەتی پڕۆسەس کردنی زماندا پێویستمان بە داشکاندنی دەق (segmentation) بۆ دەرھێنانی پێکھاتە زمانەوانییەکان یان ئاسانکاریی تەکنیکی ھەیە.

• ھەنگاویکی پێش-پڕۆسەس کردنی زۆرینە ی ئەرکەکانی پڕۆسەس کردنی زمانە

مافەکانی مڕۆف کۆمەڵێک بنچینە یان پرسیای ڕەوشتین کە لەسەر پێوانەی ئاکاریی مڕۆفەکان ساز کراون و بەگوێرەی یاسای ناوخوای و نیودەوڵەتی پارێزگارییان لێ دەکرێت. بەگشتی دەتوانین ئەم بنەمایانە بەم شێوەیە ناو بنین کە مافی ڕەوای ھەموو مڕۆفێکن و شیاوی گۆڕین و دەستکاریی نین و بەیەکسانی و بەھۆی مڕۆفبوونی کەسەکەو پێی ڕەوا ببنراو. بێ گۆیدانە تەمەن، نەتەرە، ئایین، زمان، ڕەنگ و، ڕەگەز ھەمووان وەک مڕۆف ھاوتان. ئەم مافانە لە ھەموو شوێنێک و لە ھەموو کاتیکیدا جێبەجێ دەبن بەسەر مڕۆفەکاندا، لە ھەر کۆیەک بن لە جیھاندا، بێ جیاوازی. لە سایەی سەرورەیی یاسا و سۆزی مڕۆفانەو چاوەڕوان دەکرێت کە ھەموو پێڕەوی ریزگرتنی مافەکانی مڕۆفی ھەبێت بۆ مڕۆفەکانی تر، ھێچ کات ئەو مافانە لە مڕۆف ناترازی، مافی مڕۆف بۆ ھەموو کەسێک مافیکی ڕەسەنە و پێوەی لکاوە، مەگەر لە دۆخیکی یاسایی تایبەتی و بە ھۆکاری سەپاندنی لە ڕەوتیکی یاساییدا.

وەرگیراو لە وتاری «مافەکانی مڕۆف»، ویکیپیدیای کوردی



چۆن دەق لە سەر کۆمپیوتەر نیشان دەدەین؟

لە زۆر بابەتی پڕۆسەس کردنی زماندا پێویستمان بە داشکاندنی دەق (segmentation) بۆ دەرھێنانی پێکھاتە زمانەوانییەکان یان ئاسانکاریی تەکنیکی ھەیە.

- ھەنگاویکی پێش-پڕۆسەس کردنی زۆرینە ی ئەرکەکانی پڕۆسەس کردنی زمانە
- لەتەکان یەکە ی جیاوازی دەقن (پیت، وشە، ھتد)

مافەکانی مرقوف کۆمەڵیک بنچینە یان پرسیای ڤهوشتین که لەسەر پێوانە ی ئاکاریی مرقوفەکان ساز کراون و بەگوێرە ی یاسای ناوخوا یی و نیودەولەتی پارێزگارییان لی دەکریت. بەگشتی دەتوانین ئەم بنەمایانە بەم شیۆمیە ناو بنین که مافی ڤهوا ی ھەموو مرقوفیکن و شیوا ی گوڤین و دەستکاریی نین و بەیەکسانی و بەھۆی مرقوفوونی کەسەکەو پێی ڤهوا بینراو. بێ گوێدانە تەمەن، نەتەو، ئایین، زمان، ڤهنگ و، ڤهگەز ھەمووان وەکو مرقوف ھاوتان. ئەم مافانە لە ھەموو شوینیک و لە ھەموو کاتیکیا جیبەجی دەبن بەسەر مرقوفەکاندا، لە ھەر کوێیک بن لە جیھاندا، بێ جیاوازی. لە سایە ی سەرورە یی یاسا و سۆزی مرقوفانەو ڤاوەروان دەکریت که ھەموو پێڤوی ریزگرتنی مافەکانی مرقوفی ھەبیت بۆ مرقوفەکانی تر، ھیچ کات ئەو مافانە لە مرقوف ناترازین، مافی مرقوف بۆ ھەموو کەسێک مافیکی ڤەسەنە و پێوہی لکاو، مەگەر لە ڤۆخیک یاسایی تایبەتی و بە ھۆکاری سەپاندنی لە ڤهوتیکی یاساییدا.

وەرگیراو لە وتاری «مافەکانی مرقوف»، ویکیپیدیای کوردی



چۆن دەق لە سەر کۆمپیوتەر نیشان دەدەین؟

لە زۆر بابەتی پڕۆسەس کردنی زماندا پیۆیستمان بە داشکاندنی دەق (segmentation) بۆ دەرھێنانی ییکھاتە زمانەوانییەکان یان ئاسانکاریی تەکنیکی ھەیە.

- ھەنگاوێکی پێش-پڕۆسەس کردنی زۆرینە ی ئەرکەکانی پڕۆسەس کردنی زمانە
- لەتەکان یەکە ی جیاوازی دەقن (پیت، وشە، ھتد)
- ئاستی زبر وەک داشکاندنی دەق بە سەر پاراگراف و رستەکاندا

مافەکانی مەژۆف کۆمەڵێک بنچینە یان رێسای رەوشتین کە لەسەر پێوانەی ئاکاری مەژۆفەکان ساز کراون و بەگۆڕە یاسای ناوخرۆیی و نۆدەولەتی پارێزگاریان لێ دەرکێت. بەگشتی دەتوانین ئەم بنەمایانە بەم شێوەیە ناو بنین کە مافی رەوا ی ھەموو مەژۆفێک و شیاوی گۆڕین و دەستکاریی نین و بەیەکسانی و بەھۆی مەژۆفبوونی کەسەکەو پێی رەوا ببنراو. بێ گۆیدانە تەمەن، نەتەو، ئایین، زمان، رەنگ و، رەگەز ھەمووان وەکو مەژۆف ھاوتان. ئەم مافانە لە ھەموو شوێنێک و لە ھەموو کاتێکدا جێبەجێ دەبن بەسەر مەژۆفەکاندا، لە ھەر کۆیەک بن لە جیھاندا، بێ جیاوازی. لە سایە ی سەرورە ی یاسا و سۆزی مەژۆفانەو چاومەوان دەرکێت کە ھەموو پێرە ی رێزگرتنی مافەکانی مەژۆفی ھەبێت بۆ مەژۆفەکانی تر، ھێچ کات ئەو مافانە لە مەژۆف ناترازی، مافی مەژۆف بۆ ھەموو کەسێک مافیکی رەسەنە و پێو ی لکاوە، مەگەر لە دۆخیکی یاسایی تایبەتی و بە ھۆکاری سەپاندنی لە رەویتیکی یاسایدا.

وەرگیراو لە وتاری «مافەکانی مەژۆف»، ویکیپیدیای کوردی



چۆن دهق له سهر كۆمپيوتهر نيشان دهدهين؟

له زۆر بابتهى پرۆسهس كردنى زماندا پيوستمان به داشكاندى دهق (segmentation) بو دهرهينانى پيکهاته زمانهوانيهکان يان ئاسانکاريى تهکنیکی ههيه.

- ههنگاوێكى پيش-پرۆسهس كردنى زۆرينهى ئهركهكانى پرۆسهس كردنى زمانه
- لهتهكان يهكهى جياوازی دهقن (پيت، وشه، هتد)
- ئاستى زبر وهك داشكاندى دهق به سهر پاراگراف و رستهكاندا

_ماف_مکان_ى_ مَرۆف_ کۆمهڵ_يک_ بنچينه_ يان_ ريسا_ى_ رهوشتى_ن_ که_ لهسهر_ پيوانه_ى_ ئاکارى_ى_ مَرۆف_هکان_ ساز_ کراون_ و_ بهگۆيره_ى_ ياسا_ى_ ناوخويى_ و_ نيودهولتهى_ پاريزگارى_يان_ لى_ دهکريت_.



- له تکردن یان که رتکردنی دهق چه شنیکه له داشکاندنی وشه و بهو له تانه دهلیین «توکن» (token)



- لهتکردن یان کهرتکردنی دهق چهشنیکه له داشکاندنی وشه و بهو لهتانه دهلیین «تۆکن» (token)
- بۆیه کهم جار له سالی 1992 دا ئاماژه بهو لهتانه کراوه [1]

TOKENIZATION AS THE INITIAL PHASE IN NLP

Jonathan J. Webster & Chunyu Kit
City Polytechnic of Hong Kong
83 Tat Chee Avenue, Kowloon, Hong Kong
E-mail: ctwebste@cphkvx.bitnet



- لهتکردن یان کهرتکردنی دهق چهشنیکه له داشکاندنی وشه و بهو لهتانه دهلیین «تۆکن» (token)
- بۆیه کهم جار له سالی 1992 دا ئاماژه بهو لهتانه کراوه [1]
- بنه مای زۆر ئهرکی پرۆسه سکردنی زمان له سهر تۆکنه کانه



- له تکردن یان که رتکردنی دهق چه شنیکه له داشکاندنی وشه و بهو له تانه دهلیین «توکن» (token)
- بۆ یه کهم جار له سالی 1992 دا ئاماژه بهو له تانه کراوه [1]
- بنه مای زۆر ئهرکی پرۆسه سکردنی زمان له سهر توکنه کانه

توکن

- له تکردن یان که رتکردنی دهق چه شنیکه له داشکاندنی وشه و بهو له تانه دهلیین «توکن» (token)
- بۆ یه کهم جار له سالی 1992 دا ئاماژه بهو له تانه کراوه [1]
- بنه مای زۆر ئهرکی پرۆسه سکردنی زمان له سهر توکنه کانه

توکن

- «پریز بهندی یهک له دوا ی یهک و نابه تالی گرافیم یان فونیمه کان له ده قدا» [2]

- له تکردن یان که رتکردنی دهق چه شنیکه له داشکاندنی وشه و بهو له تانه دهلیین «توکن» (token)
- بۆ یه کهم جار له سالی 1992 دا ئاماژه بهو له تانه کراوه [1]
- بنه مای زۆر ئهرکی پرۆسه سکردنی زمان له سهر توکنه کانه

توکن

- «پریز بهندی یهک له دوا ی یهک و نابه تالی گرافیم یان فوئیمه کان له ده قدا» [2]
- «له تکردنی دهق به سهر توکندا بریتییه له دۆزینه وهی سنووری وشه کان» [3]

- له تکردن یان که رتکردنی دهق چه شنیکه له داشکاندنی وشه و بهو له تانه دهلیین «توکن» (token)
- بۆ یه کهم جار له سالی 1992 دا ئاماژه بهو له تانه کراوه [1]
- بنه مای زۆر ئهرکی پرۆسه سکردنی زمان له سهر توکنه کانه

توکن

- «پریز بهندی یهک له دوا ی یهک و نابه تالی گرافیم یان فونیمه کان له ده قدا» [2]
- «له تکردنی دهق به سهر توکندا بریتییه له دۆزینه وهی سنووری وشه کان» [3]
- «توکن خه یالیکی پروپووچی ئەنداز یاریکی کو مپیوتره بۆ ئاسانکردنی کاری خو ی» [4]

لهتکردنی تۆکن: سنووری وشه

پیناسه‌ی تۆکن به پیی دۆزینه‌وه‌ی سنووری وشه

Dr. O'Neil presents state-of-the-art results. ●



له تکردنی تۆکن: سنووری وشه

پیناسه ی تۆکن به پیی دۆزینه وه ی سنووری وشه

• Dr. O'Neil presents state-of-the-art results.

• « دواکه وتنی شیوازه کانی به رهه مهینان »



له تکردنی تۆکن: سنووری وشه

پیناسه ی تۆکن به پیی دۆزینه وه ی سنووری وشه

• Dr. O'Neil presents state-of-the-art results.

• « دواکه وتنی شیوازه کانی به رهه مهینان »

• دواکه وتنی شیوازه کانی به رهه مهینان



له تکردنی تۆکن: سنووری وشه

پیناسه ی تۆکن به پیی دۆزینه وه ی سنووری وشه

Dr. O'Neil presents state-of-the-art results. •

• « دواکه وتنی شیوازهکانی به رهه مهینان »

• دواکه وتنی شیوازهکانی به رهه مهینان

• دوا کهوتنی شیوازهکانی به رهه مهینان



له تکردنی تۆکن: سنووری وشه

پیناسه ی تۆکن به پیی دۆزینه وه ی سنووری وشه

Dr. O'Neil presents state-of-the-art results. ●

● « دواکه وتنی شیوازهکانی به رهه مهینان »

● دواکه وتنی شیوازهکانی به رهه مهینان

● دوا کهوتنی شیوازهکانی به رهه مهینان

● دواکه وتنی شیوازهکانی به رهه مهینان



له تکردنی تۆکن: سنووری وشه

پیناسه ی تۆکن به پیی دۆزینه وه ی سنووری وشه

Dr. O'Neil presents state-of-the-art results. ●

● « دواکه وتنی شیواز هکانی به رهه مهینان »

● دواکه وتنی شیواز هکانی به رهه مهینان

● دوا که وتنی شیواز هکانی به رهه مهینان

● دواکه وتنی ی شیواز هکان ی به رهه مهینان

● دوا که وتنی ی شیواز هکان ی به رهه مهینان



له تکر دنی توکن: سنووری وشه

پیناسه ی توکن به پیی دۆزینه وهی سنووری وشه

Dr. O'Neil presents state-of-the-art results. •

• « دواکه وتنی شیواز هکانی به رهه مهینان »

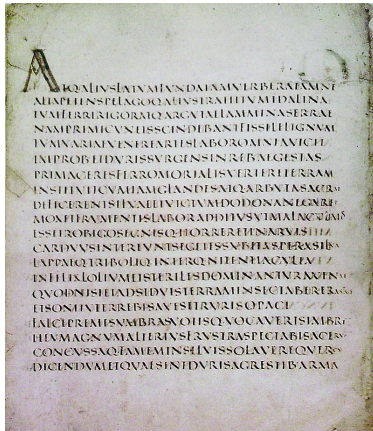
• دواکه وتنی شیواز هکانی به رهه مهینان

• دوا کهوتنی شیواز هکانی به رهه مهینان

• دواکه وتنی ی شیواز هکان ی به رهه مهینان

• دوا کهوتن ی شیواز هکان ی به رهه مهینان

• سنووری وشه له Scriptio continua هکاندا



له تکر دنی توکن: سنووری وشه

پیناسه ی توکن به پیی دۆزینه وهی سنووری وشه

Dr. O'Neil presents state-of-the-art results.

« دواکه وتنی شیواز هکانی به رهه مهینان »

• دواکه وتنی شیواز هکانی به رهه مهینان

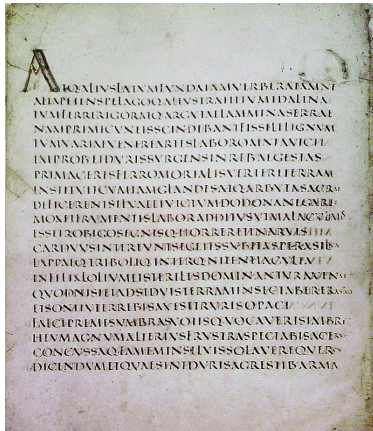
• دوا که وتنی شیواز هکانی به رهه مهینان

• دواکه وتنی ی شیواز هکان ی به رهه مهینان

• دوا که وتنی ی شیواز هکان ی به رهه مهینان

• سنووری وشه له Scriptio continua هکاندا

• سنووری وشه له زمانه کانی دیکه دا



له تکر دنی توکن: سنووری وشه

پیناسه ی توکن به پیی دۆزینه وهی سنووری وشه

Dr. O'Neil presents state-of-the-art results. •

• « دواکهوتنی شیوازهکانی به رهه مهینان »

• دواکهوتنی شیوازهکانی به رهه مهینان

• دوا کهوتنی شیوازهکانی به رهه مهینان

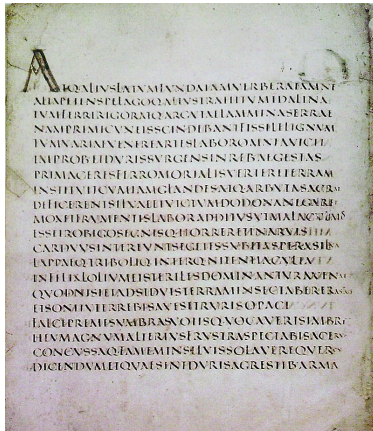
• دواکهوتن ی شیواز هکان ی به رهه مهینان

• دوا کهوتن ی شیواز هکان ی به رهه مهینان

• سنووری وشه له Scriptio continua هکاندا

• سنووری وشه له زمانهکانی دیکه دا

• 我是中国人



له تکر دنی توکن: سنووری وشه

پیناسه ی توکن به پیی دۆزینه وهی سنووری وشه

Dr. O'Neil presents state-of-the-art results. •

• « دواکه وتنی شیواز هکانی به رهه مهینان »

• دواکه وتنی شیواز هکانی به رهه مهینان

• دوا کهوتنی شیواز هکانی به رهه مهینان

• دواکه وتنی ی شیواز هکان ی به رهه مهینان

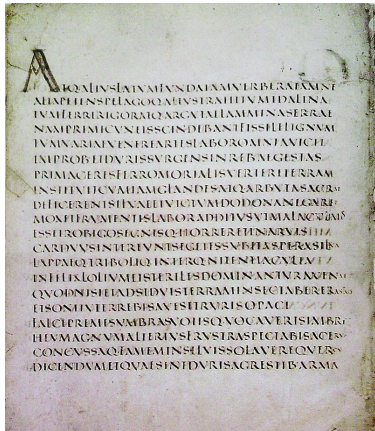
• دوا کهوتن ی شیواز هکان ی به رهه مهینان

• سنووری وشه له Scriptio continua هکاندا

• سنووری وشه له زمانه کانی دیکه دا

• 我是中国人

• つ目の入力



له تکر دنی توکن: سنووری وشه

پیناسه ی توکن به پیی دۆزینه وهی سنووری وشه

Dr. O'Neil presents state-of-the-art results. •

• « دواکهوتنی شیوازهکانی به رهه مهینان »

• دواکهوتنی شیوازهکانی به رهه مهینان

• دوا کهوتنی شیوازهکانی به رهه مهینان

• دواکهوتن ی شیواز هکان ی به رهه مهینان

• دوا کهوتن ی شیواز هکان ی به رهه مهینان

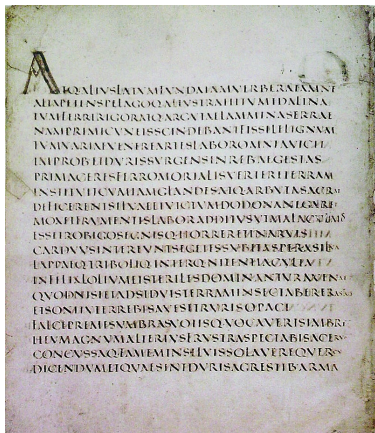
• سنووری وشه له Scriptio continua هکاندا

• سنووری وشه له زمانهکانی دیکه دا

• 我是中国人

• つ目の入力

• Sprachwissenschaft



له تکر دنی توکن: سنووری وشه

پیناسه ی توکن به پیی دۆزینه وهی سنووری وشه

Dr. O'Neil presents state-of-the-art results. •

• « دواکهوتنی شیواز هکانی به رهه مهینان »

• دواکهوتنی شیواز هکانی به رهه مهینان

• دوا کهوتنی شیواز هکانی به رهه مهینان

• دواکهوتن ی شیواز هکان ی به رهه مهینان

• دوا کهوتن ی شیواز هکان ی به رهه مهینان

• سنووری وشه له Scriptio continua هکاندا

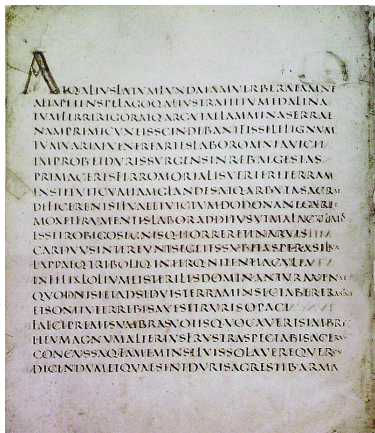
• سنووری وشه له زمانه هکانی دیکه دا

• 我是中国人

• つ目の入力

• Sprachwissenschaft

پیناسه کردنی توکن ئه رکیکی ئاسان نییه!



لهتکردنی تۆکن: روانگهی رینووس

روانگهی رینووس: پشت به شیوازی نووسینی دهقه که بیهسته



له تکردنی تۆکن: روانگهی رینووس

- روانگهی رینووس: پشت به شیوازی نووسینی دهقه که بیهسته داشکاندن به پپی هیماکانی خالبهندی وهک «!؟،:،»، بۆشایی، هیلی نوی و ته ب



له تکردنی تۆکن: روانگهی رینووس

- روانگهی رینووس: پشت به شیوازی نووسینی دهقه که بیهسته داشکاندن به پپی هیماکانی خالبهندی وهک «!؟،:،»، بوشایی، هیلی نوی و ته ب دواکه وتنی شیوازهکانی به ره مهینان .



له تکردنی تۆکن: روانگهی رینووس

- روانگهی رینووس: پشت به شیوازی نووسینی دهقه که بیهسته
- داشکاندن به پپی هیماکانی خالبهندی وهک «!؟،:،»، بوشایی، هیلی نوی و ته ب
 - دواکه وتنی شیوازهکانی به ره مهینان .
 - باشی +: ئاسانه



لهتکردنی تۆکن: روانگهی رینووس

روانگهی رینووس: پشت به شیوازی نووسینی دهقه که بیهسته

- داشکاندن به پپی هیماکانی خالبهندی وهک «!؟،:،»، بوشایی، هیلی نوی و ته ب

• دواکهوتنی شیوازهکانی بهرهمهینان

- باشی +: ئاسانه

- خراپی -



لهتکردنی تۆکن: روانگهی رینووس

روانگهی رینووس: پشت به شیوازی نووسینی دهقه که بیهسته

- داشکاندن به پپی هیماکانی خالبهندی وهک «!؟،،،» ، بوشایی، هیلی نوی و ته ب

• دواکهوتنی شیوازهکانی بهرهمهینان

- باشی +: ئاسانه

- خراپی -

- داهینانی ههموو یاساکان زهحمهته



له تکر دنی تۆکن: روانگهی رینووس

روانگهی رینووس: پشت به شیوازی نووسینی دهقه که بیهسته

• داشکاندن به پپی هیماکانی خالبهندی وهک «!؟،،،» ، بوشایی، هیلی نوی و ته ب

• دواکه وتنی شیوازهکانی به ره مهینان

• باشی +: ئاسانه

• خراپی -

• داهینانی هه موو یاساکان زه حمه ته

• پیویستی به زانیاری له سهر زمانه کان و رینووسه کانیاں هه یه



له تکر دنی تۆکن: روانگهی رینووس

روانگهی رینووس: پشت به شیوازی نووسینی دهقه که بیهسته

• داشکاندن به پپی هیماکانی خالبهندی وهک «!؟،،،»، بوشایی، هیلی نوی و ته ب

• دواکه وتنی شیوازهکانی به ره مهینان

• باشی +: ئاسانه

• خراپی -

• داهینانی هه موو یاساکان زه حمه ته

• پیویستی به زانیاری له سهر زمانه کان و رینووسه کانیاں هه یه

• ناتوانی چاره سهری سنووره دیارینه کراوهکانی وشه بکا



له تکر دنی تۆکن: روانگهی رینووس

روانگهی رینووس: پشت به شیوازی نووسینی دهقه که بیهسته

• داشکاندن به پپی هیماکانی خالبهندی وهک «!?!:;»,» بوشایی، هیلی نوی و ته ب

• دواکه وتنی شیوازهکانی به ره مهینان

• باشی +: ئاسانه

• خراپی -

• داهینانی هه موو یاساکان زه حمه ته

• پیویستی به زانیاری له سهر زمانه کان و رینووسه کانیاں هه یه

• ناتوانی چاره سهری سنووره دیارینه کراوهکانی وشه بکا

• هه لهی رینووسی سازکردنی وشه دانه که تیک ددها



له تکر دنی تۆکن: روانگهی زمانه وانی

روانگهی زمانه وانی: پۆلی زمانه وانی پیکهاته کان وهک پۆلی وشه کان له پرسته سازیدا

- له تکر دن به پیی پیکهاته کان له وشه دا، وهک مۆرفیمه کان



له تکر دنی تۆکن: روانگهی زمانه وانی

- روانگهی زمانه وانی: پۆلی زمانه وانی پیکهاته کان وهک پۆلی وشه کان له پرسته سازیدا
- له تکر دن به پیی پیکهاته کان له وشه دا، وهک مۆرفیمه کان
 - زۆر جار، تۆکنه کان به پیی سه رچاوه یکی دیکه ده رده هیئندریڻ، وهک فرههنگ



له تکر دنی تۆکن: روانگهی زمانه وانی

روانگهی زمانه وانی: رۆلی زمانه وانی پیکهاته کان وهک رۆلی وشه کان له رسته سازیدا

- له تکر دن به پیی پیکهاته کان له وشه دا، وهک مۆرفیمه کان
- زۆر جار، تۆکنه کان به پیی سه رچاوه یکی دیکه ده رده هیئندری، وهک فرههنگ

دواکه وتن ی شیوازه کان ی به رهه مهینان .



له تکر دنی تۆکن: روانگهی زمانه وانی

روانگهی زمانه وانی: رۆلی زمانه وانی پیکهاته کان وهک رۆلی وشه کان له رسته سازیدا

- له تکر دن به پیی پیکهاته کان له وشه دا، وهک مۆرفیمه کان
- زۆر جار، تۆکنه کان به پیی سه رچاوه یکی دیکه ده رده هیئندری، وهک فرههنگ

دواکه وتن ی شیوازه کان ی به رهه مهینان .

دوا که وتن ی شیواز هکان ی به رهه م هینان .



لهتکردنی تۆکن: روانگهی زمانهوانی

روانگهی زمانهوانی: پۆلی زمانهوانیی پیکهاتهکان وهک پۆلی وشهکان له پرسته سازیدا

- لهتکردن به پێی پیکهاتهکان له وشه دا، وهک مۆرفیمهکان
- زۆر جار، تۆکنهکان به پێی سه رچاوهیکی دیکه ده رده هیئندرین، وهک فرههنگ

دواکهوتن ی شیواز هکان ی به ره مهینان .

دوا کهوتن ی شیواز هکان ی به ره م هینان .

• باشی +



له تکر دنی تۆکن: روانگهی زمانه وانی

پروانگهی زمانه وانی: پۆلی زمانه وانی پیکهاته کان وهک پۆلی وشه کان له پرسته سازیدا

- له تکر دن به پیی پیکهاته کان له وشه دا، وهک مۆرفیمه کان
- زۆر جار، تۆکنه کان به پیی سه رچاوه یکی دیکه ده رده هیئندرین، وهک فرههنگ

دواکه وتن ی شیواز هکان ی به رهه مهینان .

دوا که وتن ی شیواز هکان ی به رهه م هینان .

• باشی +

• ئاسانه به لام نه به قهت پروانگهی رینووس



لهتکردنی تۆکن: روانگهی زمانهوانی

روانگهی زمانهوانی: رۆلی زمانهوانیی پیکهاتهکان وهک رۆلی وشهکان له رسته سازیدا

- لهتکردن به پێی پیکهاتهکان له وشه دا، وهک مۆرفیمهکان
- زۆر جار، تۆکنهکان به پێی سهراچاوهیکی دیکه دهردههیندرین، وهک فرههنگ

دواکهوتن ی شیوازهکان ی بهرهمهینان .

دوا کهوتن ی شیواز هکان ی بهرهم هینان .

• باشی +

- ئاسانه بهلام نه به قهت روانگهی رینووس
- بهسووده لهو ئهرکانهی که پیوستیان به بۆچوونیکی زمانهوانانه ههیه وهک لیکدانهوهی رسته



لهتکردنی تۆکن: روانگهی زمانهوانی

روانگهی زمانهوانی: رۆلی زمانهوانیی پیکهاتهکان وهک رۆلی وشهکان له رسته سازیدا

- لهتکردن به پێی پیکهاتهکان له وشه دا، وهک مۆرفیمهکان
- زۆر جار، تۆکنهکان به پێی سهراوهیکی دیکه دهردههیندرین، وهک فرههنگ

دواکهوتن ی شیوازهکان ی بهرهمهینان .

دوا کهوتن ی شیواز هکان ی بهرهم هینان .

• باشی +

• ئاسانه بهلام نه به قهت روانگهی رینووس

• بهسووده لهو ئهرکانهی که پیوستیان به بۆچوونیکی زمانهوانانه ههیه وهک لیکدانهوهی رسته

• خراپی -



له تکر دنی تۆکن: روانگهی زمانه وانی

روانگهی زمانه وانی: پۆلی زمانه وانی پیکهاته کان وهک پۆلی وشه کان له رسته سازیدا

- له تکر دن به پیی پیکهاته کان له وشه دا، وهک مۆرفیمه کان
- زۆر جار، تۆکنه کان به پیی سه رچاوه یکی دیکه ده رده هیئدرین، وهک فرههنگ

دواکه وتن ی شیواز هکان ی به رهه مهینان .

دوا که وتن ی شیواز هکان ی به رهه م هینان .

• باشی +

- ئاسانه به لام نه به قهت روانگهی رینووس
- به سووده لهو ئه رکانه ی که پیوستان به بۆچوونیکی زمانه وانانه ههیه وهک لیکدانه وهی رسته

• خراپی -

- پیوستانی به زانیاری زمانه وانی له سه ر زمانه کان و رینووسه کانیا ن ههیه



لهتکردنی تۆکن: روانگهی زمانهوانی

روانگهی زمانهوانی: رۆلی زمانهوانیی پیکهاتهکان وهک رۆلی وشهکان له پرسته سازیدا

- لهتکردن به پێی پیکهاتهکان له وشه دا، وهک مۆرفیمهکان
- زۆر جار، تۆکنهکان به پێی سه رچاوهیکی دیکه ده رده هیئدرین، وهک فرههنگ

دواکهوتن ی شیوازهکان ی به رهه مهینان .

دوا کهوتن ی شیواز هکان ی به رهه م هینان .

• باشی +

- ئاسانه به لام نه به قهت روانگهی رینووس
- به سووده لهو ئه رکانه ی که پینوستان به بۆچوونیکی زمانه وانانه ههیه وهک لیکدانه وهی پرسته

• خراپی -

- پینوستی به زانیاری زمانهوانی له سه ر زمانهکان و رینووسهکانیان ههیه
- یاسا زمانهوانیهکان دهتوانن ناتهواو بن



له تکر دنی توکن: روانگهی زمانه وانی

روانگهی زمانه وانی: رۆلی زمانه وانی پیکهاته کان وهک رۆلی وشه کان له رسته سازیدا

- له تکر دن به پیی پیکهاته کان له وشه دا، وهک مؤرفیمه کان
- زۆر جار، توکنه کان به پیی سه رچاوه یکی دیکه دهرده هیئدرین، وهک فرههنگ

دواکه وتن ی شیواز هکان ی به ره مهینان .

دوا که وتن ی شیواز هکان ی به ره م هینان .

• باشی +

- ئاسانه به لام نه به قهت روانگهی رینووس
- به سووده لهو ئه رکانه ی که پیوستان به بۆچوونیکی زمانه وانانه ههیه وهک لیکدانه وهی رسته

• خراپی -

- پیوستانی به زانیاری زمانه وانی له سه ر زمانه کان و رینووسه کانیا ههیه
- یاسا زمانه وانیه کان ده توانن ناته واو بن
- ههله ی رینووسی سازکردنی وشه دانه که تیک ده دا



له تکر دنی توکن: روانگهی زمانه وانی

روانگهی زمانه وانی: رۆلی زمانه وانی پیکهاته کان وهک رۆلی وشه کان له رسته سازیدا

- له تکر دن به پیی پیکهاته کان له وشه دا، وهک مؤرفیمه کان
- زۆر جار، توکنه کان به پیی سه رچاوه یکی دیکه دهرده هیندری، وهک فرههنگ

دواکه وتن ی شیواز هکان ی به ره مهینان .

دوا که وتن ی شیواز هکان ی به ره هم هینان .

• باشی +

- ئاسانه به لام نه به قهت روانگهی رینووس
- به سووده لهو ئه رکانه ی که پیوستیان به بۆچوونیکی زمانه وانانه ههیه وهک لیکدانه وهی رسته

• خراپی -

- پیوستی به زانیاری زمانه وانی له سه ر زمانه کان و رینووسه کانیا ههیه
- یاسا زمانه وانیه کان ده توانن ناته واو بن
- هه له ی رینووسی سازکردنی وشه دانه که تیک دهدا
- ئه گه ره وشه یه که له فرههنگدا نه بی و یاسا کان کاریان لی نه کا؟ ← وشه ی نه ناسراو



زمان پره له وشه و هه موویان له فرههنگدا ده رناکهون، وهک «تورته مالی»، «کوڤید» یان «لاتک»



- زمان پره له وشه و هه موویان له فرههنگدا ده رناکهون، وهک «تورته مالی»، «کوڤید» یان «لاتک»
- نهو تۆکنانه ی له وشه دان یان فرههنگی زمانه که دا نه هاتوون به $\langle unk \rangle$ دیاری ده کرین



له تکر دنی تۆکن: وشه نه ناسراوه کان

- زمان پره له وشه و هه موویان له فرههنگدا ده رناکهون، وهک «تورته مالی»، «کوڤید» یان «لاتک»
- ئه و تۆکنانه ی له وشه دان یان فرههنگی زمانه که دا نه هاتوون به *unk* دیاری ده کرین
 - زۆر وشه ی ده گمه ن یان نیوی که س به وشه ی نه ناسراو ده رده چن



له تکر دنی تۆکن: وشه نه ناسراوه کان

- زمان پره له وشه و هه موویان له فرههنگدا ده رناکهون، وهک «تورته مالی»، «کوڤید» یان «لاتک»
- ئه و تۆکنانه ی له وشه دان یان فرههنگی زمانه که دا نه هاتوون به $\langle unk \rangle$ دیاری ده کرین
 - زۆر وشه ی ده گمه ن یان نیوی که س به وشه ی نه ناسراو ده رده چن
- دوا کهوتن ی شیواز هکان ی $\langle unk \rangle$ هیئان .



له تکر دنی تۆکن: وشه نه ناسراوه کان

- زمان پره له وشه و هه موویان له فرههنگدا ده رناکهون، وهک «تورته مالی»، «کوڤید» یان «لاتک»
- ئه و تۆکنانه ی له وشه دان یان فرههنگی زمانه که دا نه هاتوون به *<unk>* دیاری ده کرین
 - زۆر وشه ی ده گمه ن یان نیوی که س به وشه ی نه ناسراو ده رده چن
- دوا که وتن ی شیواز هکان ی *<unk>* هیئان .
- نه ناسینه وه ی تۆکنیک ده بیته هو ی له ده ست چوونی زانیاری



له تکر دنی تۆکن: وشه نه ناسراوه کان

زمان پره له وشه و هه موویان له فرههنگدا ده رناکهون، وهک «تورته مالی»، «کوڤید» یان «لاتک»

- ئه و تۆکنانه ی له وشه دان یان فرههنگی زمانه که دا نه هاتوون به $\langle unk \rangle$ دیاری ده کرین

- زۆر وشه ی ده گمه ن یان نیوی که س به وشه ی نه ناسراو ده رده چن

دوا کهوتن ی شیواز هکان ی $\langle unk \rangle$ هینان .

- نه ناسینه وه ی تۆکنیک ده بیته هو ی له ده ست چوونی زانیاری

- چۆن ئه م کیشه یه چاره سه ر بکه ین به بی ئه وه ی پیویستمان به زانیاری له سه ر هه موو زمانه کدا هه بی؟



لهتکردنی تۆکن: روانگهی پیدراوه

روانگهی پیدراوه: چ پیکهاتهگهلیک باوتره، با به تۆکنیان دانیین
• ئەم بۆچوونه چارهسهریکه بۆ کیشهکانی بۆچوونه پیشووهکان



لهتکردنی تۆکن: روانگهی پیدراوه

روانگهی پیدراوه: چ پیکهاتهگهلیک باوتره، با به تۆکنیان دانیین

- ئەم بۆچوونه چارهسهریکه بۆ کیشهکانی بۆچوونه پیشووهکان
- پشت به فراوانیی ئەو پیتانه دههستێ که ویکرا دهردهکهون و پیاان دهلین وشهله (subword)



لهتکردنی تۆکن: روانگهی پیدراوه

روانگهی پیدراوه: چ پیکهاتهگهلیک باوتره، با به تۆکنیان دانیین

- ئەم بۆچوونه چارهسهری که بۆ کیشهکانی بۆچوونه پیشووهکان

- پشت به فراوانیی ئەو پیتانه دههستی که ویکرا دهردهکهون و پیاان دهلین وشهله (subword)

- **Byte pair encoding (BPE)**: ناسراوترین و باوترین ئەلگوریتمه (algorithm) که پشت بهم بۆچوونهوه دههستی [5]



ههنگاو هکانی راهینان:

1 سازکردنی وشه دان: دهست پیده کات به دابه شکردنی وشه کان بو تاکه پیته کان (هه ریه کیکیان هیمایه که له وشه دانی کو تاییدا)



ههنگاوهکانی راهینان:

1 سازکردنی وشه دان: دهست پیدهکات به دابهشکردنی وشهکان بۆ تاکه پیتهکان (ههر یهکیکیان هیمایه که له وشه دانی کووتاییدا)

2 دۆزینهوهی جووته هیمای فراوان له وشه دانی ئیستادا. بهم ئهرکه دهلین یهکخستن (merge)

بۆ نموونه ئهگهر پیتی «د» به فراوانیهکی زۆرهوه له گهڵ «ه»
دهرکهوی، «ده» وهکوو یهک هیما له وشه دانه که زیاد دهکا



ههنگاههکانی راهینان:

- 1 سازکردنی وشه دان: دهست پیدهکات به دابهشکردنی وشهکان بو تاکه پیتهکان (ههر یهکیکیان هیمایه که له وشه دانی کو تاییدا)
- 2 دوزینه وهی جووته هیمای فراوان له وشه دانی ئیستادا. بهم ئهرکه دهلین یهکخستن (merge)
بو نموونه ئهگهر پیتی «د» به فراوانیهکی زوره وه له گهل «ه»
دهرکه وئی، «ده» وهکوو یهک هیما له وشه دانه که زیاد دهکا
- 3 زیادکردنی هیما یهکخراوهکان له وشه دانه که



ههنگاوهکانی راهینان:

- 1 سازکردنی وشه دان: دهست پیدهکات به دابهشکردنی وشهکان بۆ تاکه پیتهکان (ههر یهکیکیان هیمایه که له وشه دانی کو تاییدا)
- 2 دۆزینهوهی جووته هیمای فراوان له وشه دانی ئیستادا. بهم ئهرکه دهلین یهکخستن (merge)
بۆ نموونه ئهگهر پیتی «د» به فراوانییهکی زۆره وه له گهڵ «ه»
دهرکهوئ، «ده» وهکوو یهک هیما له وشه دانه که زیاد دهکا
- 3 زیادکردنی هیما یهکخراوهکان له وشه دانه که
- 4 دووباره کردنهوهی ههنگاوهکانی (2) و (3) ههتا به ژمارهی دیاریکراوی یهکخران دهگهین یان هیچ پیکهاتهیهکی نویی هیماکان به فراوانی پێویسته وه نه مینیت



لهتکردنی تۆکن: ریکاری BPE - نموونه

low low low low low lowest lowest newer newer
newer newer newer newer wider wider wider new new



له تکر دنی توکن: ریکاری BPE - نمونه

```
low low low low low lowest lowest newer newer  
newer newer newer newer wider wider wider new new
```

1 سازکردنی وشه دان: دابه شکردنی وشه کان بو تاکه پیته کان. دهقه که به پیی بو شایی داشکینه و کو تایی وشه کان به هیمایه کی تایبته دیاری که وهک «</w>» یان «_»

کو رپس

```
'l o w </w>': 5  
'l o w e s t </w>': 2  
'n e w e r </w>': 6  
'w i d e r </w>': 3  
'n e w </w>': 2
```



له تکرندی توکن: ریکاری BPE - نمونه

```
low low low low low lowest lowest newer newer  
newer newer newer newer wider wider wider new new
```

1 سازکردنی وشه دان: دابه شکردنی وشه کان بو تاکه پیته کان. دهقه که به پیی بو شایی داشکینه و کوتاییی وشه کان به هیمایه کی تایبته دیاری که وهک «</w>» یان «_»

کۆرپس

```
'l o w </w>': 5  
'l o w e s t </w>': 2  
'n e w e r </w>': 6  
'w i d e r </w>': 3  
'n e w </w>': 2
```

وشه دان

```
d, e, i, l, n, o, r, s, t, w, </w>
```



لهتکردنی تۆکن: ریکاری BPE - نموونه

1 سازکردنی وشه‌دان

کۆرپس

'l o w </w>': 5

'l o w e s t </w>': 2

'n e w e r </w>': 6

'w i d e r </w>': 3

'n e w </w>': 2

وشه‌دان

d, e, i, l, n, o, r, s, t, w, </w>



لهتکردنی تۆکن: ریکاری BPE - نموونه

فراوانیی جووتهکان

('e', 'r'): 9
('r', '</w>'): 9
('w', 'e'): 8
('n', 'e'): 8
('e', 'w'): 8
('l', 'o'): 7
('o', 'w'): 7
('w', '</w>'): 7
('w', 'i'): 3
('i', 'd'): 3
('d', 'e'): 3
('e', 's'): 2
('s', 't'): 2
('t', '</w>'): 2

1 سازکردنی وشه‌دان

کۆرپس

'l o w </w>': 5
'l o w e s t </w>': 2
'n e w e r </w>': 6
'w i d e r </w>': 3
'n e w </w>': 2

وشه‌دان

d, e, i, l, n, o, r, s, t, w, </w>

2 یه‌کخستنی جووته فراوانه‌کان

لهتکردنی تۆکن: ریکاری BPE - نموونه

فراوانیی جووتهکان

('e', 'r'): 9
('r', '</w>'): 9
('w', 'e'): 8
('n', 'e'): 8
('e', 'w'): 8
('l', 'o'): 7
('o', 'w'): 7
('w', '</w>'): 7
('w', 'i'): 3
('i', 'd'): 3
('d', 'e'): 3
('e', 's'): 2
('s', 't'): 2
('t', '</w>'): 2

1 سازکردنی وشه‌دان

کۆرپس
'l o w </w>': 5
'l o w e s t </w>': 2
'n e w e r </w>': 6
'w i d e r </w>': 3
'n e w </w>': 2

وشه‌دان

d, e, i, l, n, o, r, s, t, w, </w>

2 یه‌کخستنی جووته فراوانه‌کان

'e', 'r' ده‌بن به یه‌ک وشه‌له \Leftarrow er

له تکر دنی توکن: ریکاری BPE - نمونه

- 1 سازکردنی وشه دان
- 2 یه کخستنی جووته فراوانه کان
- 3 زیادکردنی هیما یه کخراوه کان له وشه دان هه

کۆرپس

'l o w </w>': 5

'l o w e s t </w>': 2

'n e w e r </w>': 6

'w i d e r </w>': 3

'n e w </w>': 2

وشه دان

d, e, i, l, n, o, r, s, t, w, </w>, er



لهتکردنی تۆکن: ریکاری BPE - نموونه

فراوانیی جووتهکان

('er', '</w>'): 9
('n', 'e'): 8
('e', 'w'): 8
('l', 'o'): 7
('o', 'w'): 7
('w', '</w>'): 7
('w', 'er'): 6
('w', 'i'): 3
('i', 'd'): 3
('d', 'er'): 3
('w', 'e'): 2
('e', 's'): 2
('s', 't'): 2
('t', '</w>'): 2

1 سازکردنی وشه‌دان

2 یه‌کخستنی جووته فراوانه‌کان

3 زیادکردنی هیما یه‌کخراوه‌کان له وشه‌دانه‌که

کورپس

'l o w </w>': 5

'l o w e s t </w>': 2

'n e w e r </w>': 6

'w i d e r </w>': 3

'n e w </w>': 2

وشه‌دان

d, e, i, l, n, o, r, s, t, w, </w>, er

لهتکردنی تۆکن: ریکاری BPE - نموونه

فراوانیی جووتهکان

('er', '</w>'): 9
('n', 'e'): 8
('e', 'w'): 8
('l', 'o'): 7
('o', 'w'): 7
('w', '</w>'): 7
('w', 'er'): 6
('w', 'i'): 3
('i', 'd'): 3
('d', 'er'): 3
('w', 'e'): 2
('e', 's'): 2
('s', 't'): 2
('t', '</w>'): 2

1 سازکردنی وشه‌دان

2 یه‌کخستنی جووته فراوانه‌کان

3 زیادکردنی هیما یه‌کخراوه‌کان له وشه‌دانه‌که

کۆرپس
'l o w </w>': 5
'l o w e s t </w>': 2
'n e w e r </w>': 6
'w i d e r </w>': 3
'n e w </w>': 2

وشه‌دان
d, e, i, l, n, o, r, s, t, w, </w>, er

4 هه‌نگاوی 2 و 3 دووپات بکه‌وه

لهتکردنی تۆکن: ریکاری BPE - نمونه

ههنگاوی 2 و 3 دووپات بکهوه

فراوانیی جووتهکان

('n', 'e'): 8
('e', 'w'): 8
('l', 'o'): 7
('o', 'w'): 7
('w', '</w>'): 7
('w', 'er</w>'): 6
('w', 'i'): 3
('i', 'd'): 3
('d', 'er</w>'): 3
('w', 'e'): 2
('e', 's'): 2
('s', 't'): 2
('t', '</w>'): 2

$er</w> \Leftarrow 'er', '</w>'$

کۆرپس

'l o w </w>': 5
'l o w e s t </w>': 2
'n e w er</w>': 6
'w i d er</w>': 3
'n e w </w>': 2

وشه‌دان

d, e, i, l, n, o, r, s, t,
w, </w>, er, er</w>



لهتکردنی توکن: ریکاری BPE - نمونه

ههنگاوی 2 و 3 دوو پات بکهوه

فراوانیی جووتهکان

('ne', 'w'): 8
('l', 'o': 7
('o', 'w': 7
('w', '</w>': 7
('w', 'er</w>': 6
('w', 'i': 3
('i', 'd': 3
('d', 'er</w>': 3
('w', 'e': 2
('e', 's': 2
('s', 't': 2
('t', '</w>': 2

ne \leftarrow 'n', 'e'

کۆرپس

'l o w </w>': 5
'l o w e s t </w>': 2
'ne w er</w>': 6
'w i d er</w>': 3
'ne w </w>': 2

وشه دان

d, e, i, l, n, o, r, s, t,
w, </w>, er</w>, er, ne

له تکرندی توکن: ریکاری BPE - نمونه

ریکاری BPE له پاش 10 یه کخستن دهگا بهمانه:

فراوانیی جووتهکان

```
('wid', 'er</w>'): 3  
( 'low', 'e'): 2  
( 'e', 's'): 2  
( 's', 't'): 2  
( 't', '</w>'): 2  
( 'new', '</w>'): 2
```

کورپس

```
'low</w>': 5  
'low e s t </w>': 2  
'newer</w>': 6  
'wider</w>': 3  
'new </w>': 2
```

وشه‌دان d, e, i, l, n, o, r, s, t, w, </w>
er, er</w>, ne, new, lo,
low, newer</w>, low</w>, wi, wid



له تکرڊنی تۆکن: ریکاری BPE

- له کاردا، زۆربهی مۆدیلهکان و ئامیرهکان 32 بۆ 64 ههزار وشه له یان له وشه داندا ههیه بۆ نمونه، 50257 GPT2 وشه له ی ههیه



له تکردنی توکن: ریکاری BPE

- له کاردا، زۆربهی مۆدیلهکان و ئامیژهکان 32 بۆ 64 ههزار وشه له یان له وشه داندا ههیه بۆ نموونه، 50257 GPT2 وشه له ی ههیه
- وشه له کان دهتوانن مۆرفیمه کان بن وهک «یک» یان وشه باوهکان وهک «ههیه» یان ههر پیکهاتهیهکی باو



له تکردنی توکن: ریکاری BPE

- له کاردا، زۆربهی مۆدیلهکان و ئامیژهکان 32 بۆ 64 ههزار وشه له یان له وشه داندا ههیه بۆ نموونه، 50257 GPT2 وشه له ی ههیه
- وشه له کان ده توانن مۆرفیمه کان بن وهک «یک» یان وشه باوه کان وهک «ههیه» یان ههر پیکهاتهیه کی باو
- له پاش ئه ژماریکی دیاریکراوی یه کخستن له سهر کوژیسیکی گه وره، وشه دانه که بۆ له تکردنی توکنه کان به کار دی



له تکرډنی تۆکن: ریکاری BPE

- له کاردا، زۆر به ی مۆدیلله کان و ئامپیره کان 32 بۆ 64 ههزار وشه له یان له وشه داندا ههیه بۆ نموونه، 50257 GPT2 وشه له ی ههیه
- وشه له کان ده توانن مۆرفیمه کان بن وهک «یک» یان وشه باوه کان وهک «ههیه» یان ههر پیکهاتهیه کی باو
- له پاش ئه ژماریکی دیاریکراوی یه کخستن له سهر کۆرپسیکی گه وره، وشه دانکه بۆ له تکرډنی تۆکنه کان به کار دی
- له تکرډنی تۆکنه کان بۆ ده قیکی نوی به پپی ئه و وشه لانه ده کری که فیریان بووین کهوا بی، گرینگ نییه دهقه نوییه که چی تیدایه. وشه دانکه ته نیا پشت به کۆرپسه که ده به ستی.



له تکرندی تۆکن: ریکاری BPE

- له کاردا، زۆر به ی مۆدیلەکان و ئامیڕەکان 32 بۆ 64 ههزار وشە له یان له وشە داندا ههیه بۆ نموونه، 50257 GPT2 وشە له ی ههیه
- وشە له کان ده توانن مۆرفیمه کان بن وهک «یک» یان وشە باوه کان وهک «ههیه» یان ههر پیکهاته یه کی باو
- له پاش ئەژماریکێ دیاریکراوی یه کخستن له سهر کۆرپسیکی گه وره، وشه دانکه بۆ له تکرندی تۆکنه کان به کار دی
- له تکرندی تۆکنه کان بۆ ده قیکی نوی به پپی ئەو وشە لانه ده کری که فیریان بووین کهوا بی، گرینگ نییه دهقه نوییه که چی تیدا یه. وشه دانکه ته نیا پشت به کۆرپسه که ده به ستی.
- ئەم ریکاره له زۆر مۆدیل و ئامیڕدا به کار دی، به تایبەت له مۆدیلەکانی زمان (language models)



له تکردنی تۆکن: ریکاری BPE

- له کاردا، زۆر به ی مۆدیلەکان و ئامیڕەکان 32 بۆ 64 هەزار وشە لە یان له وشە داندا هەیه بۆ نموونه، 50257 GPT2 وشە لە ی هەیه
- وشە لە کان دەتوانن مۆرفیمەکان بن وهک «یک» یان وشە باوهکان وهک «ههیه» یان ههر پیکهاتهیهکی باو
- له پاش ئەژماریکێ دیا ریکراوی یه کخستن له سهر کۆرپسیکی گهوره، وشه دانکه بۆ له تکردنی تۆکنهکان به کار دی
- له تکردنی تۆکنهکان بۆ دهقیکی نوێ به پێی ئەو وشە لانه دهکری که فیریان بووین کهوا بی، گرینگ نییه دهقه نوییه که چی تیدایه. وشه دانکه ته نیا پشت به کۆرپسه که ده به ستی.
- ئەم ریکاره له زۆر مۆدیل و ئامیڕدا به کار دی، به تایبەت له مۆدیلەکانی زمان (language models)
- BPE ته نیا ریکار له سهر وشه له کان نییه
[6] Unigram و [7] WordPiece



له‌تکردنی تۆکن و ده‌برپینی فره‌وشه

- ده‌برپینی فره‌وشه (multiword expression) پیکهاتوو له چهند به‌ش که ویکرا واتایه‌کی جودایان هه‌یه



له تکرندی تۆکن و ده‌برپینی فره‌وشه

- ده‌برپینی فره‌وشه (multiword expression) پیکهاتوو له چهند به‌ش که ویکرا واتایه‌کی جودایان هه‌یه

• برپاریان دا ← برپار یان دا



له تکردنی تۆکن و ده‌برپینی فره‌وشه

- ده‌برپینی فره‌وشه (multiword expression) پیکهاتوو له چهند به‌ش که ویکرا واتایه‌کی جودایان هه‌یه

• برپاریان دا ← برپار یان دا

• خشکه و پشکه ← خشکه و پشکه



له تکردنی تۆکن و ده‌برپینی فره‌وشه

- ده‌برپینی فره‌وشه (multiword expression) پیکهاتوو له چه‌ند به‌ش که وی‌کرا واتایه‌کی جودایان هه‌یه

• برپاریان دا ← برپار یان دا

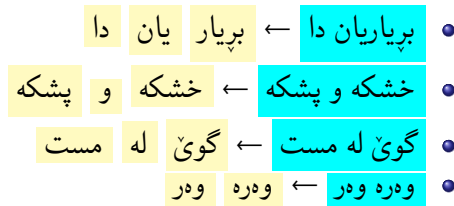
• خشکه و پشکه ← خشکه و پشکه

• گوێ له مست ← گوێ له مست



له تکر دنی توکن و دهر برینی فره وشه

- دهر برینی فره وشه (multiword expression) پیکهاتوو له چهند بهش که ویکرا واتایهکی جودایان ههیه



لهتکردنی تۆکن و دهربرینی فرهوشه

- دهربرینی فرهوشه (multiword expression) پیکهاتوو له چهند بهش که ویکرا واتایهکی جودایان ههیه

•	←	بپاریان دا	بپار	یان	دا
•	←	خشکه و پشکه	خشکه	و	پشکه
•	←	گوی له مست	گوی	له	مست
•	←	وهه وهه	وهه	وهه	
•	←	bi-can-û-bên	bên	û	can bi



لهتکردنی تۆکن و دهربرینی فرهوشه

- دهربرینی فرهوشه (multiword expression) پیکهاتوو له چهند بهش که ویکرا واتایهکی جودایان ههیه

• بریاریان دا ← بریار یان دا

• خشکه و پشکه ← خشکه و پشکه

• گوئی له مست ← گوئی له مست

• وهره وهر ← وهره وهر

• bi-can-û-bên ← bi can û bên

- چۆن ئهم دهربرینانه لهت بکرین؟



لهتکردنی تۆکن و دهربرینی فرهوشه

- دهربرینی فرهوشه (multiword expression) پیکهاتوو له چهند بهش که ویکرا واتایهکی جودایان ههیه

• بپاریان دا ← بپار یان دا

• خشکه و پشکه ← خشکه و پشکه

• گوئی له مست ← گوئی له مست

• وهره وهر ← وهره وهر

• bi-can-û-bên ← bi can û bên

- چۆن ئهم دهربرینانه لهت بکرین؟

- بهکارهینانی فرههنگی تایبته به دهربرینه فرهوشهیهکان



لهتکردنی تۆکن و دهربرینی فرهوشه

- دهربرینی فرهوشه (multiword expression) پیکهاتوو له چهند بهش که ویکرا واتایهکی جودایان ههیه

• بپاریان دا ← بپار یان دا

• خشکه و پشکه ← خشکه و پشکه

• گوئی له مست ← گوئی له مست

• وهره وهر ← وهره وهر

• bi-can-û-bên ← bi can û bên

- چۆن ئهم دهربرینانه لهت بکرین؟

- بهکارهینانی فرههنگی تایبته به دهربرینه فرهوشهیهکان
- بهکارهینانی روانگهی پیدراوه به رهچاوکردنی بۆشایی؟



- مۆدیله نۆراله کان (neural models) ی زمان سیسته مگه لی ته نیشته به ته نیشته (end-to-end) ن له پرۆسه س کردنی زماندا و جیگه ی رپیره وه نه ریتییه کانیا ن گرتوو ته وه.



له تکر دنی تۆکن: چه مکی ئیستا

- مۆدیله نۆراله کان (neural models) ی زمان سیسته مگه لی ته نیشته به ته نیشته (end-to-end) ن له پرۆسه س کردنی زماندا و جیگه ی رپیره وه نه ریتییه کانیا ن گرتوو ته وه.
- چه مکی ئیستای له تکر دنی تۆکن: پشتبهستن به پیدراوه له باتی یاسای زمانه وانی



- مۆدیله نۆراله کان (neural models) ی زمان سیسته مگه لی ته نیشته به ته نیشته (end-to-end) ن له پرۆسه س کردنی زماندا و جیگه ی رپرۆه وه نه ریتییه کانیا ن گرتو وه ته وه.
- چه مکی ئیستای له تکردنی تۆکن: پشتبهستن به پیدراوه له باتی یاسای زمانه وانی
- دابه شکردن بۆ یه که ی بچوو کتر و ناتایپوگرافی وه ک وشه له کان

لهتکردنی تۆکن: چه مکی ئیستا

- مۆدیله نۆرالهکان (neural models) ی زمان سیسته مگه لی ته نیشته به ته نیشته (end-to-end) ن له پرۆسه س کردنی زماندا و جیگه ی رپیره وه نه ریتییه کانیا ن گرتوو ته وه.
- چه مکی ئیستای لهتکردنی تۆکن: پشتبهستن به پیدراوه له باتی یاسای زمانهوانی
 - دابهشکردن بۆ یه که ی بچوو کتر و ناتایپوگرافی وهک وشه لهکان
 - تۆکنهکان له گه ل رینووس یان یه که ی زمانهوانیدا یهک ناگر نه وه \Leftarrow تۆکن \neq وشه



له تکر دنی تۆکن: چه مکی ئیستا

- مۆدیله نۆراله کان (neural models) ی زمان سیسته مگه لی ته نیشته به ته نیشته (end-to-end) ن له پرۆسه س کردنی زماندا و جیگه ی رپرته وه نه ریتییه کانیا ن گرتو وه ته وه.
- چه مکی ئیستای له تکر دنی تۆکن: پشت به ستن به پیدرا وه له باتی یاسای زمانه وانی
 - دابه شکردن بۆ یه که ی بچوو کتر و ناتایپو گرافی وه ک وشه له کان
 - تۆکنه کان له گه ل رینووس یان یه که ی زمانه وانیدا یه ک ناگر نه وه \Leftarrow تۆکن \neq وشه
 - له تکر دنی تۆکنه کان به پپی رینووس یان زمانه وانی \Leftarrow «له تکر دنی پیشه کیی» (pre-tokenization)



له تکر دنی تۆکن: چه مکی ئیستا

- مۆدیله نۆراله کان (neural models) ی زمان سیسته مگه لی ته نیشته به ته نیشته (end-to-end) ن له پرۆسه س کردنی زماندا و جیگه ی پریره وه نه ریتییه کانیا ن گرتوو ته وه.
- چه مکی ئیستای له تکر دنی تۆکن: پشتبه ستن به پیدراوه له باتی یاسای زمانه وانی
 - دابه شکردن بۆ یه که ی بچوو کتر و ناتایپوگرافی وه ک وشه له کان
 - تۆکنه کان له گه ل رینووس یان یه که ی زمانه وانیدا یه ک ناگر نه وه \Leftarrow تۆکن \neq وشه
 - له تکر دنی تۆکنه کان به پپی رینووس یان زمانه وانی \Leftarrow «له تکر دنی پی شه کیی» (pre-tokenization)
 - تۆکنه نه ریتییه کان \Leftarrow «پیش-تۆکن» (pre-token)



- «یه‌ک به‌ش که به که‌لله‌ره‌قی له کۆنه‌وه ماوه‌ته‌وه لهتکردنی تۆکنه» [8]

لهتکردنی تۆکن و مۆدیله گهورهکانی زمان

- «یهک بهش که به کهله رهقی له کۆنه وه ماوه ته وه لهتکردنی تۆکنه» [8]
- «شکۆی ههتا ههتایی بۆ هه رکه سییک که بتوانی لهتکردن وهک ههنگاویکی پیویست له مۆدیله زمانییه گهورهکاندا (large language models, LLMs) بسریتته وه» (Andrej Karpathy)



لهتکردنی تۆکن بۆ کوردی



له نووسینی کوردیدا، بوڤشایی و نیشانهی نیو بوڤشایی (ZWNJ) بو جودا کردنه وهی وشه کان به کار دین، به لام کیشهی زۆریان ههیه:

- یه کلانه بوونه وهی رینووس:



له نووسینی کوردیدا، بوڤشایی و نیشانهی نیو بوڤشایی (ZWNJ) بو جودا کردنه وهی وشه کان به کار دین، به لام کیښه ی زۆریان ههیه:

• یه کلانه بوونه وهی رینووس:

• 18ê, 18-ê, 18'ê



له نووسینی کوردیدا، بۆشایی و نیشانهی نیوبۆشایی (ZWNJ) بۆ جوداکردنهوهی وشهکان بهکار دێن، بهلام کیشهی زۆریان ههیه:

• یهکلانهبوونهوهی پینووس:

• 18ê ,18-ê ,18'ê

• بهکار دێ، بهکار دێ، بهکار دێ، بهکار دێ



له نووسینی کوردیدا، بۆشایی و نیشانهی نیوبۆشایی (ZWNJ) بۆ جوداکردنهوهی وشهکان بهکار دیڤن، بهلام کیشهی زۆریان ههیه:

- یهکلانهبوونهوهی پینووس:

- 18ê, 18-ê, 18'ê

- بهکار دی، بهکار دی، بهکار دی، بهکار دی، بهکار دی

- نووساندنی زۆری وشهکان:



له نووسینی کوردیدا، بۆشایی و نیشانهی نیو بۆشایی (ZWNJ) بۆ جوداکردنهوهی وشهکان بهکار دێن، بهلام کیشهی زۆریان ههیه:

- یهکلانهبوونهوهی پینووس:

- 18ê ,18-ê ,18'ê

- بهکار دێ، بهکار دێ، به کار دێ، به کار دێ

- نووساندنی زۆری وشهکان:

- لهویشدايه ← له وی ش دا یه



له نووسینی کوردیدا، بۆشایی و نیشانهی نیو بۆشایی (ZWJ) بۆ جوداکردنهوهی وشهکان بهکار دێن، بهلام کیشهی زۆریان ههیه:

- یهکلانهبوونهوهی رینووس:

- 18ê, 18-ê, 18'ê

- بهکار دێ، بهکار دێ، بهکار دێ، بهکار دێ

- نووساندنی زۆری وشهکان:

- لهویشدايه ← له وی ش دا یه

- وشه لیکدراوهکان:

مردوو مرو

خواخراو بۆکردگ

«له کوردستانهوه کووتایی به پێوهندییهکه هیئا.» ← له ... هوه؟ کووتایی هیئا؟



له تکردنی تۆکن بو کوردی له KLPT دا

- له تکردنی تۆکن له KLPT دا پشت به فهرههنگیکی وشه و شیکاریی مۆرفۆلۆژی ده بهستی



له تکردنی تۆکن بۆ کوردی له KLPT دا

- له تکردنی تۆکن له KLPT دا پشت به فهرههنگیکی وشه و شیکاریی مۆرفۆلۆژی ده بهستی
- گونجاوه بۆ له تکردنی زمانهوانی



له تکردنی تۆکن بۆ کوردی له KLPT دا

- له تکردنی تۆکن له KLPT دا پشت به فهرههنگیکی وشه و شیکاریی مۆرفۆلۆژی ده بهستی
- گونجاوه بۆ له تکردنی زمانهوانی
- یارمهتی دۆزینهوهی سنووری وشه ده دا



له تکردنی تۆکن بۆ کوردی له KLPT دا

- له تکردنی تۆکن له KLPT دا پشت به فهرههنگیکی وشه و شیکاریی مۆرفۆلۆژی ده به ستن
- گونجاوه بۆ له تکردنی زمانهوانی
- یارمهتی دۆزینهوهی سنووری وشه ده دا
- ریخۆشکهره بۆ ئهرکی دیکه له زمانهوانیی کۆمپیوتهریدا



له تکرڊنی تۆکن بۆ کوردی له KLPT دا

- له تکرڊنی تۆکن له KLPT دا پشت به فهرههنگیکی وشه و شیکاریی مۆرفۆلۆژی ده به ستن
- گونجاوه بۆ له تکرڊنی زمانه وانی
- یارمه تی دۆزینه وهی سنووری وشه ده دا
- ریخۆشکهره بۆ ئهرکی دیکه له زمانه وانی کۆمپیوته ریدا
- نمونه ی له تکرڊن:

دهق: به رهه مهینانی شیوازهکانی دواکهوتنی
دوای له تکرڊن: دوا-کهوتن-ی شیوازهکان-ی به رهه م-هینان-ی



لهتکردنی تۆکن بۆ کوردی له KLPT دا

- لهتکردنی تۆکن له KLPT دا پشت به فهرههنگیکی وشه و شیکاریی مۆرفۆلۆژی ده‌به‌ستێ
- گونجاوه بۆ لهتکردنی زمانه‌وانی
- یارمه‌تی دۆزینه‌وه‌ی سنووری وشه‌ ده‌دا
- ریخۆشکه‌ره بۆ ئه‌رکی دیکه له زمانه‌وانیی کۆمپیوته‌ریدا
- نمونه‌ی لهتکردن:

ده‌ق: به‌ره‌مه‌ییانی شیوازه‌کانی دواکه‌وتنی
دوای لهتکردن: دوا-که‌وتن-ی شیوازه‌کان-ی به‌ره‌م-هیئان-ی

- له‌م وتاره‌دا باسی کراوه: [10]



لهتکردنی تۆکن بۆ کوردی له KLPT دا

- لهتکردنی تۆکن له KLPT دا پشت به فهرههنگیکی وشه و شیکاریی مۆرفۆلۆژی ده‌به‌ستێ
- گونجاوه بۆ لهتکردنی زمانه‌وانی
- یارمه‌تی دۆزینه‌وه‌ی سنووری وشه‌ ده‌دا
- ریخۆشکهره‌ بۆ ئهرکی دیکه‌ له‌ زمانه‌وانیی کۆمپیوته‌ریدا
- نمونه‌ی له‌تکردن:

ده‌ق: به‌ره‌مه‌ییانی شیوازه‌کانی دواکه‌وتنی
دوای له‌تکردن: دوا-که‌وتن-ی شیوازه‌کان-ی به‌ره‌م-هیئان-ی

- له‌م وتاره‌دا باسی کراوه: [10]

• KLPT: <https://github.com/sinaahmadi/klpt>



1 کۆ کردنهوهی وشه



1 کۆ کردنهوهی وشه

2 وشه لیکدراوهکان بدۆژهوه

کوردیی ناوهندی: 8180 وشه (1513 لیکدراو)
کوردیی سهروو: 9970 وشه (1507 لیکدراو)



- 1 کۆ کردنه وهی وشه
- 2 وشه لیکدراوه کان بدۆزه وه
کوردیی ناوهندی: 8180 وشه (1513 لیکدراو)
کوردیی سهروو: 9970 وشه (1507 لیکدراو)
- 3 هه موو فۆرمه ئالۆز و نائالۆزه کانیا ن ساز بکه



نموونه‌ی کوردیی سه‌روو

"bi-can-û-bên":

"bicanûbên"

"bi canûbên"

"bican ûbên"

"bi can ûbên"

"bicanû bên"

"bi canû bên"

1 کۆ کردنه‌وه‌ی وشه

2 وشه لی‌کدراوه‌کان بدۆزه‌وه

کوردیی ناوه‌ندی: 8180 وشه (1513 لی‌کدراو)
کوردیی سه‌روو: 9970 وشه (1507 لی‌کدراو)

3 هه‌موو فۆرمه ئالۆز و نائالۆزه‌کانیان ساز بکه



نموونه‌ی کوردیی سه‌روو

"bi-can-û-bên":

"bicanûbên"

"bi canûbên"

"bican ûbên"

"bi can ûbên"

"bicanû bên"

"bi canû bên"

1 کۆ کردنه‌وه‌ی وشه

2 وشه لێکدراوه‌کان بدۆژه‌وه
کوردیی ناوه‌ندی: 8180 وشه (1513 لێکدراو)
کوردیی سه‌روو: 9970 وشه (1507 لێکدراو)

3 هه‌موو فۆرمه ئالۆز و نا‌ئالۆزه‌کانیان ساز بکه

نموونه‌ی کوردیی ناوه‌ندی

"ئاخر-و-ئۆخر":

"ئاخرو ئۆخر"

"ئاخرووئۆخر"

"ئاخر و ئۆخر"

"ئاخرووئۆخر"

"ئاخر وئۆخر"



- لیستی پێشگر، پاشگر و کلیتیکهکان:
پێشگر: له، وه، ده، ره، به
پاشگر: هکه، هکان، ان، گهل
جیناوه لکاوهکان: م، ت، ی، مان، تان، یان



- لیستی پێشگر، پاشگر و کلیتی که کان:
پێشگر: له، وه، ده، ره، به
پاشگر: هکه، هکان، ان، گهل
جیناوه لکاوهکان: م، ت، ی، مان، تان، یان
- نموونهی شیکاریی مۆرفۆلۆژی:

بهرزترهکه ← بهرز + تر + هکه

- لیستی پێشگر، پاشگر و کلیتی که کان:
پێشگر: له، وه، ده، ره، به
پاشگر: هکه، هکان، ان، گهل
جیناوه لکاوهکان: م، ت، ی، مان، تان، یان
- نموونهی شیکاریی مۆرفۆلۆژی:

بهرزترهکه ← بهرز + تر + هکه

- ژمارهی مۆرفیمهکان:
سۆرانی: 161 پاشگر و 11 پێشگر
کورمانجی: 46 پاشگر و 17 پێشگر



1

پیش-پروژهس کردن:

یه کخستنی نووسینی پیتەکان
زیادکردنی بۆشایی له دهوری خالبهندی و ژمارهکان

نموونه

« ٢٥ کهسیان پێ کهوتن»
1. 25 کهسیان پێ کهوتن



نموونه

«٢٥ کهسیان پێی کهوتن»

1. 25 کهسیان پێی کهوتن

2. 25 کهسیان —پێی— کهوتن—

1 پێش-پروژهس کردن:

یه‌ک‌خستنی نووسینی پێته‌کان

زیادکردنی بۆشایی له ده‌وری خاڵبه‌ندی و ژماره‌کان

2 له‌تکردنی وشه‌ لی‌ک‌دراوه‌کان:

دۆزینه‌وه‌ی وشه‌ لی‌ک‌دراوه‌کان له‌ فهره‌ه‌نگدا

جودا‌کردنه‌وه‌یان به‌ نیشانه‌ی ژیره‌یل –

نمونه

«۲۵ کهسیان ری کهوتن»

1. 25 کهسیان ری کهوتن

2. 25 کهسیان ری-کهوتن

3. 25 کهسیان ری-کهوتن

1 پیش-پرۆسهس کردن:

یهکخستنی نووسینی پیتەکان

زیادکردنی بۆشایی له دهوری خالبەندی و ژمارەکان

2 لهتکردنی وشە لیکدراوەکان:

دۆزینەوێ وشە لیکدراوەکان له فەرھەنگدا

جوداکردنەوێان بە نیشانەیی ژیرھیل –

3 لهتکردنی وشە:

جیاکردنەوێ بە بۆشایی

دۆزینەوێ وشەکان له فەرھەنگدا

نموونه

«٢٥ کهسیان پێ کهوتن»

1. 25 کهسیان پێ کهوتن

2. 25 کهسیان —پێ— کهوتن—

3. 25 —کهسیان— —پێ— کهوتن—

4. 25 —کهس— —یان— —پێ— کهوتن—

1 پێش- پرۆسەس کردن:

یەكخستنی نووسینی پیتەكان

زیادکردنی بۆشایی لە دەوری خاڵبەندی و ژمارەكان

2 لەتکردنی وشە لیكدراوەكان:

دۆزینه‌وه‌ی وشە لیكدراوەكان لە فەرھەنگدا

جوداکردنە‌وه‌یان بە نیشانە‌ی ژێرھێڵ –

3 لەتکردنی وشە:

جیاکردنە‌وه‌ بە بۆشایی

دۆزینه‌وه‌ی وشەكان لە فەرھەنگدا

4 شیکاریی مۆرفۆلۆژی:

دۆزینه‌وه‌ی پێشگر و پاشگرەكان

جیاکردنە‌وه‌ی مۆرفیمە لیكدراوەكان



- قه باره‌ی بچووکی فرهه‌نگ:
کاریگه‌ری له سه‌ر کوالیته‌ی له‌تکردن



- قەبارەیی بچووکی فەرەهەنگ:

کاریگەری لە سەر کوالیتی لەتکردن

- ئالۆز بوونی واتای پیکهاتهکان:

بۆ نموونه وشەیی «لاوین» که دهکری «لاو» + «ین» یان وهک یهک وشه بێ، بهلام سیستمه که به
'-لاوین' لهتی دهکا



- قه باره ی بچووکی فرههنگ:
کاریگه ری له سه رکوالیتی له تکردن
- ئالۆز بوونی واتای پیکهاته کان:
بۆ نموونه وشه ی «لاوین» که دهکری «لاو» + «ین» یان وهک یهک وشه بی، به لام سیسته مه که به
'_لاوین' له تی دهکا
- پیکهاته په رژ و بلاوه کان:
وهک له «برپاریکی گه وره یان دا» دا، هه رچه ند «برپار دان» کرداریکی لیکدراوه به لام وشه ی دیکه یان
که وتوو ده ته نیوان

- قەبارە ی بچووکی فەرھەنگ:
کاریگەری لە سەر کوالیتی لەتکردن
- ئالۆز بوونی واتای پیکهاتهکان:
بۆ نموونه وشە ی «لاوین» که دهکری «لاو» + «ین» یان وهک یهک وشه بێ، بهلام سیستمه که به
'_لاوین' له تی دهکا
- پیکهاته پەرژ و بلاوهکان:
وهک له «بریاریکی گهوره یان دا» دا، هه رچه ند «بریار دان» کرداریکی لیکدراوه به لام وشه ی دیکه یان
که وتوووته نیوان
• کیشە ی کارایی:



- قەبارە ی بچووکی فەرھەنگ:
- کاریگەری لە سەر کوالیتی لەتکردن
- ئالۆز بوونی واتای پیکهاتهکان:
- بۆ نموونه وشە ی «لاوین» که دهکری «لاو» + «ین» یان وهک یهک وشه بی، بهلام سیستمه که به
'-لاوین' لهتی دهکا
- پیکهاته پەرژ و بلاوهکان:
- وهک له «بریاریکی گهورهیان دا» دا، ههچهند «بریار دان» کرداریکی لیکدراوه بهلام وشه ی دیکهیان
که وتوووته نیوان
- کیشە ی کارایی:
- گه‌ران له فەرھەنگ کاتگیره



- قەبارە ی بچووکی فەرھەنگ:
- کاریگەری لە سەر کوالیتی لەتکردن
- ئالۆز بوونی واتای پیکهاتهکان:
بۆ نموونە وشە ی «لاوین» که دهکری «لاو» + «ین» یان وهک یهک وشه بی، بهلام سیستمه که به
'_لاوین' له تی دهکا
- پیکهاته پەرژ و بلاوهکان:
وهک له «بریاریکی گهوره یان دا» دا، هه رچه ند «بریار دان» کرداریکی لیکدراوه به لام وشه ی دیکه یان
که وتوووته نیوان
- کیشە ی کارایی:
• گه ران له فەرھەنگ کاتگیره
• روانگە ی پیدراوه؟



هه‌سه‌نگاندنی له‌تکردنی تۆکن

«دواکه‌وتنی شیواز هکانی به‌ره‌مه‌یینان له‌م ئابووریانه‌دا ده‌گه‌رپێته‌وه‌ بۆ: نه‌بوونی هۆیه‌کانی ته‌کنیکی تازهی هاورده‌ تا به‌ره‌مه‌یینه‌کان به‌کاری بیّن.»

(4000) BPE

دواکه‌وت نی شیواز هکانی به‌ره‌مه‌یینان
له‌م ئابووریانه‌دا ده‌گه‌رپێته‌وه‌ بۆ: نه‌بوونی
هۆیه‌کانی ته‌کنیکی تازهی هاورده
ه‌ تا به‌ره‌مه‌یین هکان به‌کاری بیّن .

KLPT

دواکه‌وتنی شیواز هکانی به‌ره‌مه‌یینان
له‌م ئابووریانه‌دا ده‌گه‌رپێته‌وه‌ بۆ: نه‌بوونی
هۆیه‌کانی ته‌کنیکی تازهی هاورده
تا به‌ره‌مه‌یینه‌کان به‌کاری بیّن .

هه‌سه‌نگه‌ندی له‌تکردنی تۆکن پێوه‌ندی به‌ ئه‌رکه‌وه‌ هه‌یه‌.



سەرچاوەی هەموو وێنەکان ویکیمیדיا: https://commons.wikimedia.org/wiki/Main_Page

- [1] Jonathan J. Webster and Chunyu Kit.
Tokenization as the initial phase in NLP.
In *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*, 1992.
- [2] Eric de la Clergerie and Lionel Clément.
MAF: a morphosyntactic annotation framework.
Actes de LTC, pages 90–94, 2005.
- [3] David D Palmer.
Tokenisation and sentence segmentation.
Handbook of natural language processing, pages 11–35, 2000.
- [4] Sabrina J Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y Lee, Benoît Sagot, et al.
Between words and characters: a brief history of open-vocabulary modeling and tokenization in NLP.
arXiv preprint arXiv:2112.10508, 2021.
- [5] Rico Sennrich, Barry Haddow, and Alexandra Birch.
Neural machine translation of rare words with subword units.
In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.



- [6] Taku Kudo.
Subword regularization: Improving neural network translation models with multiple subword candidates.
arXiv preprint arXiv:1804.10959, 2018.
- [7] Mike Schuster and Kaisuke Nakajima.
Japanese and Korean voice search.
In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE, 2012.
- [8] Kris Cao and Laura Rimell.
You should evaluate your language model on marginal likelihood over tokenisations.
arXiv preprint arXiv:2109.02550, 2021.
- [9] Aleksandar Petrov, Emanuele La Malfa, Philip HS Torr, and Adel Bibi.
Language model tokenizers introduce unfairness between languages.
arXiv preprint arXiv:2305.15425, 2023.
- [10] Sina Ahmadi.
A tokenization system for the Kurdish language.
In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 114–127, 2020.

